

تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية

مباحث لغوية ٦٠

تحرير:

د. يوسف سالم العريان

تأليف:

د. يوسف سالم العريان

د. أمجد يوسف أبوجبارة

د. عرفان أحمد

د. أحمد الحايك

د. أحمد حمدي أبو عبسة

أ. غريب واجب غريبي

مباحث لغوية ٦٠

تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية

تأليف:

د.يوسف سالم العريان
د.أمجد يوسف أبوجبارة
د.عرفان أحمد
د.أحمد الحايك
د.أحمد حمدي أبو عبسة
أ.غريب واجب غريبي

تحرير:
د.يوسف سالم العريان

١٤٤١هـ - ٢٠١٩م





تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية

الطبعة الأولى
١٤٤١ هـ - ٢٠١٩ م

المملكة العربية السعودية - الرياض
ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ١٤٤١هـ.
فهرسة مكتبة الملك فهد الوطنية أثناء النشر
العريان، يوسف
تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية./ يوسف العريان.- الرياض، ١٤٤٠هـ
..ص؛ .. سم
ردمك: ٧ -٥٨ -٨٢٢١ -٦٠٣ -٩٧٨
١ - الذكاء الاصطناعي أ. العنوان
ديوي ٠٠٦,٣ ١١٣٠٤/ ١٤٤٠
رقم الإيداع: ١١٣٠٤/ ١٤٤٠
ردمك: ٧ -٥٨ -٨٢٢١ -٦٠٣ -٩٧٨

التصميم والإخراج
دار وجوه للنشر والتوزيع
Wojooh Publishing & Distribution House
WOJOOOH
www.wojoooh.com
المملكة العربية السعودية - الرياض
الهاتف:4562410 الفاكس:4561675
للتواصل والنشر:
info@wojoooh.com



هذا المشروع

مشروع تأليف سلسلة كتب في مجال (حوسبة العربية) يهدف إلى بناء تراكمٍ معرفي في مجال حيوي مهم، هو مجال (حوسبة العربية) . ويعد هذا الكتاب واحدا من سلسلة كتب صدرت في المركز.

يقع هذا المشروع ضمن سلسلة (مباحث لغوية) التي يشرف المركز على اختيار عنواناتها، وتكليف المحررين والمؤلفين، ومتابعة التأليف حتى إصدار الكتب. وهي سلسلة يجتهد المركز أن تكون سداداً لحاجات بحثية وعلمية تحتاج إلى تنبيه الباحثين عليها، أو تكثيف البحث فيها.

ويعدّ هذا الكتاب واحداً من كتب ثلاثة مترابطة في مشروع علمي واحد متخصص في (الذكاء الاصطناعي) :

١. العربية والذكاء الاصطناعي.
٢. تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية.
٣. خوارزميات الذكاء الاصطناعي في تحليل النص العربي.

مدير مشروع (العربية والذكاء الاصطناعي)

د.عبدالله بن يحيى الفيفي

كلمة المركز

يعمل المركز في مجال البحث العلمي ونشر الكتب مستهدفاً التركيز على المجالات البحثية التي ما زالت بحاجة إلى تسليط الضوء عليها، وتكثيف البحث فيها، ولفت أنظار الباحثين والجهات الأكاديمية إلى أهمية استثمارها بمختلف وجوه الاستثمار، وذلك مثل مجال (التخطيط اللغوي) و (العربية في العالم) و(الأدلة والمعلومات) و (تعليم العربية لأبنائها أو لغير الناطقين بها) إلى غير ذلك من المجالات، وإن من أهم مجالات البحث المستقبلية في اللغة العربية مجال (العربية والحوسبة ، والذكاء الاصطناعي) حيث إن اللغات الحية مرهونة حياتها مستقبلا بمدى تجاوبها مع التطورات التقنية والعالم الافتراضي، وكثافة المحتوى الإلكتروني المكتوب، وهو ما يشكّل تحديا حقيقيا أمام اللغات غير المنتجة للمعرفة أو للتقنية.

وقد عمل المركز على تسليط الضوء على هذا المجال التخصصي؛ مستعينا بالكفاءات القادرة من المهتمين بالتخصص البيني (بين اللغة والحاسوب) مقدّرا جهودهم، وهادفاً إلى نشرها، وتعميم مبادئها، راغباً أن يكون هذا المسار العلمي مقررا في الجامعات في كلية العربية والحاسوب، ومجالا بحثيا يقصده الباحثون الأكديميون، والجهات البحثية العربية.

وقد أصدر المركز سابقا ستة عشر كتاباً مختصا في (حوسبة العربية) وفي الإفادة من (المدونات اللغوية) في الأبحاث العربية، ويحتفل بإصدار سبعة كتب جديدة مختصة في (حوسبة العربية والذكاء الاصطناعي)، ويقدمها للقارئ العربي، وللجهات الأكاديمية؛ للإفادة منها واعتماد ما تراه منها مناسباً لتعليمه والبناء عليه، وهذه الكتب السبعة هي: (العربية والذكاء الاصطناعي، تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية، خوارزميات الذكاء الاصطناعي في تحليل النص العربي ، مقدمة في حوسبة اللغة العربية، الموارد اللغوية الحاسوبية، المعالجة الآلية للنصوص العربية، تطبيقات أساسية في المعالجة الآلية للغة العربية).

ويشكر المركز السادة مؤلفي الكتب، ومحرريها، لما تفضلوا به من عمل علمي رصين، وأدعو الباحثين والمؤلفين إلى التواصل مع المركز لاستكمال المسيرة، وتفتيق فضاءات المعرفة.

وفق الله الجهود وسدد الرؤى.

الأمين العام

أ. د. محمود إسماعيل صالح

تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية

مقدمة المحرر[1]

الحمد لله، علم الإنسان ما لم يعلم: قلما، وبيانا، وقرآنا، وخَلقا. والصلاة والسلام على النبي الأمي الذي أُرسل للعالمين سراجا منيرا. وبعد، فالذكاء الاصطناعي يُتيح وَكْلَ بعض مهام البشر للآلات، وفي بؤرته: تأليل معالجة اللغات. واللغة العربية فذة، لها فلسفات عظمى في رسمها، وفي لفظها، وفي صرفها وإعرابها وبلاغتها. لذلك تظافرت أبحاث اللغويين والحاسوبيين –عرباً وعجماً– وتسابقت للغوص عن مكنوناتها وحِكَمها، ولكنهم –للأسف– قصروا عن الانتهاء بجهودهم إلى تطبيقات عملية تصل ليد المستخدم العربي –أفرادا أو مؤسسات–، إذ كانت أكثر الجهود متفرقة، والأهداف متشعثة.

١- د.يوسف سالم العريان باحث في الحوسبة العربية، حصل على درجة الدكتوراه في علوم وهندسة الحاسب الآلي عن رسالته في «تحليل وتصنيع الكتابة العربية» من جامعة الملك فهد للبترول والمعادن، وعلى درجة الماجستير في هندسة الحاسب الآلي عن رسالته في «إنتاج معجم لعملية التعرف الآلي على الكتابة العربية» من جامعة العلوم والتكنولوجيا الأردنية. حرر كتاب «الحرف العربي والتقنية» وله العديد من الأبحاث وبراءات الاختراع في المجال. عمل محاضرا في جامعة الملك فهد للبترول والمعادن أثناء دراسته، ثم أستاذا مساعدا في جامعة جازان، ثم مدربَ تعلم وتدريس في المدينة المنورة. حائز على عدة جوائز للتميز في التدريس الأكاديمي والبحث العلمي.

لذا، فقد ارتأينا ترتيب شيء من هذا النتاج الغزير وتركيزه في بوتقة واحدة، وجعلناها عربية كي يَفيد منها الجميع: اللغوي، والحاسوبي، وغيرهم. وبذلنا -جميعا- موسوعنا في تعريب المصطلحات وأسماء المخترعات، وتقريبها للقارئ العربي (مع إبقاء أصلها بالإنجليزية ليسْهل رجوع المهتم لها في مصادرها)، وذلك بعد أن لمسنا -التقصير في التعريب الرصين ونشره، وغرابة وقع بعض الترجمات حتى على المختص.

جاء الكتاب في خمسة أبواب، تناولت قراءة الكتابة العربية آليا، والاستماع لأحكام التلاوة القرآنية تلقائيا، واستخراج الآراء والمشاعر من النصوص إلكترونياً. وقد وجدنا الباحثين قد أجمعوا -على اختلاف مشاربهم- على أهمية تقنية التعلم العميق وعلو كعبها؛ فجاء الباب الرابع ليشرح هذه التقنية. وناسب هذا كله ختم الكتاب بتطبيق لتوليد النصوص العربية الشعرية باستخدم تلكم التقنية.

فبدأ الكتاب بالتعرف والتحليل، وانتهى بالإنشاء والتطبيق، كأنه يصعد بالقارئ من الأساس إلى ذروة السنام، نسأل الله أن ينفعنا -كاتبيه وقارئيه- به. ولعل المستقبل يسفر عن كتاب يبدأ حيث انتهى هذا، يتناول ما وصل إليه العلم في تقليد لغة الإنسان، بالخط الشبيه باليدوي، والنطق العربي الطبعي، وتحليل وإنشاءٍ وتلخيص لا نكاد نفرقه عن البشري. كما أرجو أن تكون الجهود المباركة -ولعل أهمها جهود مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية- سببا لاستخلاص التطبيقات العملية من الجهود العلمية، وجني ثمارها في الدارين، والله العلي على كل شيء قدير.

وكتبه،

د. يوسف سالم العريان

ذو القعدة ١٤٤٠ هـ

عناوين أبواب الكتاب

الباب الأول

القراءة الآلية لكتابة اليد العربية

د. يوسف العريان و د. عرفان أحمد

القراءة الآلية لكتابة اليد العربية

د. يوسف العريان و د. عرفان أحمد[1]

ملخص

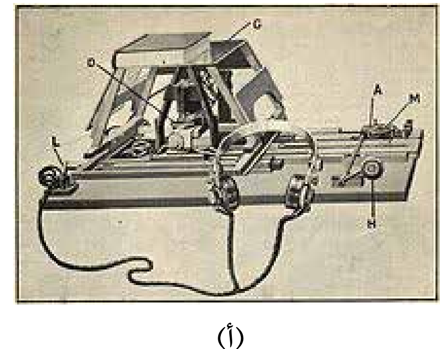
يهدف هذا الباب للأخذ بيد القارئ من مقدمات القراءة الآلية إلى أحدث تطورات مجالها. فبعد التطرق لأهم خصائص الكتابة العربية، يعرض الباب أهم عمليات التعرف الآلي على الكتابة اليدوية من معالجة مسبقة وتقطيع لحروف واستخراج ملامح وتصنيف ومعالجة لاحقة، مع التركيز على المقاربات المختلفة لمعضلة تقطيع النص العربي إلى محارفه تقطيعا صريحا أو ضمنيا أو كليا.

يشرع المؤلفان بعد ذلك بتبيان أحدث البحوث -وخاصة ما يستعمل مصنفات نماذج ماركوف الخفية والتعلم العميق- ويعرضان نتائجها ويعقدان المقارنات بينها بعد تمهيد ذلك بشرح أهم قواعد البيانات المشتهرة في تقرير نسب نجاح التعرف الآلي على الكتابة العربية اليدوية. وفي ختام الباب فصلٌ للتعريف بأبرز المجلات والمؤتمرات ذات العلاقة، لتساعد المهتم في الرجوع إلى أمهات البحوث في مظانها وليعرف أهم بواتق النشر المتاحة.

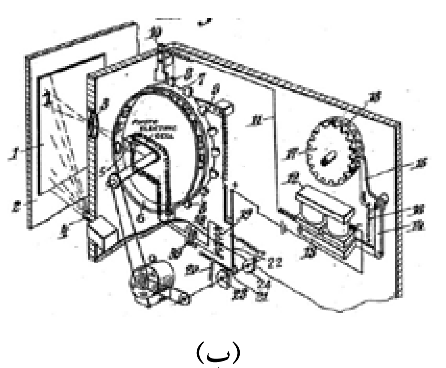
١- مقدمة

راود حلم «القارئ الآلي» البعضَ حتى قبل ظهور الحاسبات الآلية نفسها؛ ولا أدل على ذلك من تسجيل براءات اختراع لقارئات آلية ميكانيكية (كالتي في شكل ١) وتصنيع بواكيرها في القرن التاسع عشر [١،٢]. هدفت هذه الاختراعات في البداية لمساعدة ذوي الاحتياجات الخاصة بصرياً، ثم وجدت طريقها لتطبيقات أخرى كتوزيع البريد وإحصاء السكان [٢] ورقمنة الكتب والمخطوطات [٣].

١- د.عرفان أحمد أستاذ مساعد في قسم علوم الحاسب الآلي والمعلومات بجامعة الملك فهد للبترول والمعادن. حصل على درجة الدكتوراه في التعرف على الأنماط وتعلم الآلة من جامعة دورتموند التقنية بألمانيا وعلى درجة الماجستير في علوم الحاسب الآلي من جامعة الملك فهد للبترول والمعادن. نشر العديد من البحوث في مجلات ومؤتمرات مهمة، كما نشر بابا في كتاب وله عدة براءات اختراع أمريكية.



(أ)



(ب)

شكل (١): (أ) جهاز الأوبتوفون الذي يحول حروف الكتابة إلى نغمات و(ب) جزء من توصيف «الآلة القارئة» في براءة اختراع.

ومع ظهور الحاسبات الآلية، بدأت محاولات برمجتها للقراءة الآلية (أو التعرف الضوئي\البصري على النصوص Optical Character Recognition)؛ وذلك لأن تحويل صور الكتابات إلى نصوص حاسوبية (Text) يسهل معالجاتها كالبحث فيها وتحريرها وتخزينها ونقلها. شكل ٢ يعرض صورتين ونصيهما ويتيح المقارنة بين مساحاتهم التخزينية ويبرز إمكانية البحث والتحرير في النصوص.

يتم نسخ هذه الصفحة لتجرب ضمن بحث لاحق لدراسة كيف نقوم بتمييز الحروف العربية آلياً سواء كانت مكتوبة أو مطبوعة طباعة. تتضمن هذه التجربة كتابة النص من قبل عدد كبير من الناس مع ضرورة وجود بعض الإختلافات في المستوى العلمي للمشاركين مع شيء من التنوع في الأعمار ثم بعد ذلك يتم إدخال صور هذه النصوص إلى برنامج أو نظام حاسوبي وظيفته مقارنتها مع نفس النص المحفوظ أصلاً في ذاكرة الحاسوب ومن ثم استخراج واستنباط الصفات أو المميزات التي تؤدي إلى تمييز المقاطع والحروف. يفترض هنا في هذا النص الوضوح وأن يشتمل على كل حروف لغة الضاد وأن يظل صحيح اللغة. أرجو أن تحاول كتابة الكلمات التالية بالرغم من غرابتها: محمد، الحج، الكرك، صاغ، ضوؤه، إشراق، تعبوي، ثلاث مثلثات، الاكتظاظ، استنساخ، الجيش، يتلألأ، الحائط، صائغ، الخباز، بئر، شآم، يتأمل، لا تحسب ما يلي!؟؟ "((١+٢-٣)÷٤\٥) = (٦+٧-٨٠)*٩-س"	يتم نسخ هذه الصفحة لتجرب ضمن بحث لاحق لدراسة كيف نقوم بتمييز الحروف العربية آليا، سواء كانت مكتوبة أو مطبوعة طباعة. تتضمن هذه التجربة كتابة النص من قبل عدد كبير من الناس مع ضرورة وجود بعض الإختلاف في المستوى العلمي للمشاركين مع شيء من التنوع في الأعمار ثم بعد ذلك يتم ادخال صور هذه النصوص الى برنامج أو نظام حاسوبي وظيفته مقارنتها مع نفس النص المحفوظ أصلا في ذاكرة الحاسب ومن ثم استخراج واستنباط الصفات أو المميزات التي تؤدي إلى تمييز المقاطع والحروف. يفترض في هذا النص الوضوح وأن يشتمل على كل حروف لغة الضاد وأن يظل صحيح اللغة. أرجو أن تحول كتابة الكلمات التالية بالرغم من غرابتها: محمد، الحج، الكرك، صاغ، ضوؤه، اشراق، تعبوي، ثلاث مثلثات، الاكتظاظ، استنساخ، الجيش، يتلألأ، الحائط، صائغ، الخباز، بئر، شآم، يتأمل، لا تحسب ما يلي!؟؟ «((١+٢-٣)÷٤\٥) = (٦+٧-٨٠)*٩- س“
حجم الملف: ١,٠٧ ميغا بايت (١,١٢٦,٤٠٠ بايت)	حجم الملف: ١٢ كيلو بايت (١٢,٢٨٨ بايت)
امتداد الملف: BMP موحد اللون (أبيض وأسود)	امتداد الملف: DOCX

بلغ حاج أن أخاه ظمآن بوادي عوف. طفق يسعى لإحضار ثلاث قرب زمزم تنجيه مع سطوع وهيج الشمس. حث عوض الشيخ نوح بصدد ذلك فأكرمه وصب وتكلف وقال للآت أعظم ضبط سهيل وأشخاص لص الحي. غش راجح غشامة لذا جن بغيظ وانقض. انتهت.	بلغ حاج أن اخاه ظمآن بوادي عوف. طفق يسعى لإحضار ثلاث قرب زمزم تنجيه مع سطوع وهيج الشمس. حث عوض الشيخ نوح بصدد ذلك فأكرمه وصب وتكلف وقال للآت أعظم. ضبط سهيل وأشخاص لص الحي. غش راجح غشامة لذا جن بغيظ وانقض. انتهت.
حجم الملف: ٢٨٤ كيلو بايت (٢٩٠,٨١٦ بايت)	حجم الملف: ٢١١ بايت
امتداد الملف: TIF ملون	امتداد الملف: TXT

شكل (٢): أمثلة بيانات حاسوبية صورية ونصية [٤] و[٥].

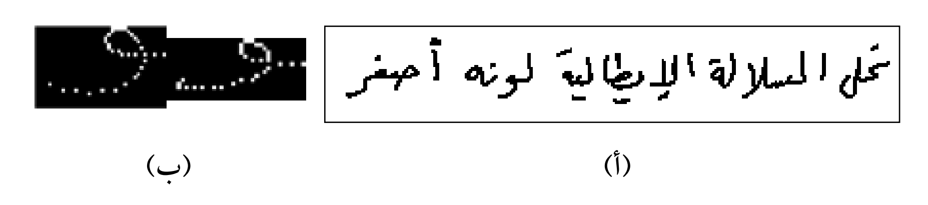
والتعرف الآلي على الكتابة من مجالات الذكاء الاصطناعي، والتي تهدف -عموما- لمحاكاة بعض قدرات البشر، ومنها التعرف على الأنماط وتمييز الحروف، بيد إن مجال القراءة الآلية أصبح يتضمن أيضا عمليات مصاحبة من مجالات كمعالجة الصور ولسانيات الحاسب الآلي، كتحديد مواضع الكتابة في الصور، وتحسين جودة الصور لتسهيل التعرف على كتابتها، وتصحيح نتائج التعرف الآلي على الكتابة لغويا.

١,١ أقسام القارئات الآلية

تقسّم أكثر التصنيفات الحديثة المتعرفات الآلية من حيث نوع المدخلات إلى نوعين:

- التعرف على التراخي (أو المنفصل (offline)) والذي يتعرف على الكتابة الورقية المكتوبة سالفا
- والتعرف الآني (أو المتصل (online)) والذي يتم أثناء الكتابة على لوحات لمس (Tablets).

وقد يُظَن من الاسمين أن التعرف المتراخي أسهل من الآني لأنه لا يتطلب سرعة الإنجاز لمواكبة عملية الكتابة في الوقت الحقيقي (Real Time)، لكن الحقيقة-وخاصة مع تسارع المعالجات- أن الكفة ربما تتجه لنجاح التعرف الآني، وذلك لتوفر بيانات لا تتوفر في الأوراق له، كترتيب رسم الحروف وأجزائها الزمني، وسرعة خطها، ومدى ضغط القلم، وكذلك لعدم تشوشه بنوع القلم وسمكه كما في الكتابة الورقية. (شكل ٣ يوضح تمثيل الكتابة اللوحية بخط موحدالسماكة وبالنقاط، حيث يمثل تباعد النقاط سرعة الكتابة).



(ب) (أ)

شكل (٣): (أ) الكتابة الآنية و(ب) تمثيل البيانات الزمنية [٥][٦].

وقد تصنف المتعرفات الآلية أيضا حسب طبيعة الكتابة والصور التي تستهدفها، كالتعرف على الخط المطبوع (ولا يكون إلا على التراخي) وخط اليد (ويمكن أن يكون على التراخي كما يمكن أن يكون آنيا). كما قد تصنف المتعرفات على التراخي حسب مصدر الصورة (من «الماسحات» (scanners)، أو من الصور الطبيعية (الناتجة من آلات التصوير أو «الكاميرات»، أو حتى من المقاطع المرئية أو «الفيديو»).

ويمكن تقسيم المتعرفات التي تستهدف الكتابة الموصولة (Cursive Writing) كالعربية إلى متعرفات تسعى لتقطيع النصوص إلى حروفها أولا، أو للتعرف على الكلمات كليا (دون تقطيعها مسبقا إلى حروف)، أو فيما يسمى بالتقطيع الضمني.

كما يمكن تقسيم المتعرفات حسب تطبيقاتها، والتي منها: رقمنة المخطوطات [٣]، وقراءة لوحات السيارات، ومعالجة السندات المصرفية (الشيكات) [٧]، وتوزيع طرود البريد، وتفريغ الاستبيانات آليا، والتعرف على كلمات اللافتات في الصور الطبيعية [٢،٨].

٢, ١ أهم تحديات التعرف الآلي على الكتابة العربية اليدوية (خط اليد العربي)

ثمة تحديات قد تواجه المشتغلين في التعرف على خط اليد -عموما-، كتغير رسم الحروف بين الكتّاب أو حتى للكاتب نفسه في مواضع وأوقات مختلفة، وخاصة إذا تغيرت الحالة النفسية أو سرعة الكتابة أو وضعيتها ومكانها وسطحها وقلمها. فهذه تحديات تظهر في خط اليد للكتابات العربية واللاتينية والصينية؛ غير أن لكل كتابة تحديات خاصة بها، لذا سنذكر في النقاط التالية بعض تحديات التعرف الآلي على الكتابة العربية:

- تغير شكل الحروف العربية المنفصلة عن تلك التي تأتي متصلة بما قبلها أو بما بعدها أو بهما معا (قارن -مثلا- أشكال حرف العين «ع» و«ـعـ» و«عـ» و«ـع»). وسنصطلح على تسمية أشكال الحروف المختلفة حسب موضعها بـ«المحارف» (Character-Shapes).
- استعمال النَّقْط لتمييز بعض الحروف المتشابهة في أصلها، ومعرفة مواضع النُّقَط من الحروف وأعدادها. ويزيد الأمر تعقيدا في الكتابة اليدوية، حيث قد يُتساهل برسمها قبيل أو بعيد الحرف وبتنوع زائد في أشكالها بناء على الخط الذي يختاره الكاتب (لاحظ النقط في شكل ٤).
- التشكيل وهو اختياري، مما يجعل للكلمة الواحدة أشكالا كثيرة صحيحة، مما قد يعقد عمل المتعرفات خاصة مع تشابه بعض النقط مع بعض التشكيل حجما وموضعا ورسما.
- إمكان التراكب الرأسي لكثير من الحروف العربية المتجاورة عوضا عن التوالي الأفقي [٩].

(جـ)	(ب)	(أ)
ثم	ثُمَّ	ثم

شكل (٤): كلمة «ثم» (أ) بدون تشكيل ولا تراكب و(ب) بتشكيلين و(جـ) بالتراكب الرأسي والنقط المتصل.

- إنفصال رسم الكلمات عند ورود حروف لا تتصل بما بعدها (أي حروف الألف والدال والذال والراء والزاي والواو ومهموزاتها وممدوداتها) أثناء الكلمة، فلا الكلمات تأتي دائما متصلة ولا الحروف تكون كلها منفصلة. ومن ذلك أيضا الانفصال عند ورود الهمزة المتطرفة على السطر بعد حرف ساكن كما في «دفْء»، و«شيء»، إذا تمنع قواعد الإملاء اتصال الحرف قبل الأخير بها وإن كان في أصله يتصل بما بعده.

- كثرة أشكال الكلمات العربية (إذا ما عرفت الكلمة بأنها ما يفصل بالمسافات وعلامات الترقيم) بسبب اللواصق السابقة (مثل «باء الجر»، و«لام التعريف» التي تتصل بأول الكلمة أو مثل «واو العطف» و«ألف الاستفهام» التي قد ترد في أوائل الكلمات لكن دون اتصال) واللواصق اللاحقة (مثل «تاء التأنيث» و«واو الجماعة») والدواخل (كما في جموع التكسير). فمثلا، كلمة «باب» في اللغة الإنجليزية هي (**door**). وهي نفسها تظهر في عبارة (“and the **door**”) بينما تظهر مختلفة بسبب السوابق الملتصقة بها في عبارة («**والباب**») [١٠].

ولكن في المقابل، فاللغة العربية تتمتع بخاصية قد تسهل قراءتها (والتعرف عليها آليا)، وهي أن لوصل الحروف وفصلها قواعد لا يجوز الحيد عنها لا طباعة ولا خطا، وهذا بخلاف الكتابة اللاتينية المعاصرة -مثلا- حيث لا يمكننا التنبؤ بما سيصله الكاتب من حروفها وما لن يصله، وهو مما قد يزيد التعرف على تلك الكتابات غموضا وصعوبة عن العربية، وهو ما توضحه أمثلة شكل ٥.

	(أ)	(ب)
الكلمة المطبوعة	الصفحة	Meeting
الكلمة المخطوطة	الصفحة	meeting

شكل (٥): مثالان يوضحان (أ) توحد طرق اتصال الحروف في الكلمة العربية و(ب) واختلافها في الحروف اللاتينية [١١][٤].

ولعل هذا ما حدا ببعض الباحثين الغربيين لأن يقول: إن العربية أسهل وأوضح اللغات في العالم، ومهما اقترحت تسهيلها وتوضيحها لم يمكن ذلك. ولو استلمت أي رسالة -مهما كانت مسطورة بخط سيء- فلن تواجه صعوبة في قراءتها [٢٤].

ونختم مقدمة الباب بذكر ترتيب فصوله الباقية، حيث يتناول الفصل الثاني عمليات التعرف الآلي على الكتابة -عموما-. أما الفصل الثالث، فيفصّل الطرق المختلفة لهيكلة عمليتي تقطيع النصوص مع التعرف عليها، فيما يُخصص الفصل الرابع للتعريف

بأشهر تجميعات الكتابة اليدوية العربية التي تستعمل في اختبار المتعرفات الآلية وتقرير نتائجها والمقارنة عبرها بين نتائج أهم أبحاث المجال. بعدهما نتمم فائدة الباب بفصل يسرد أهم أوعية النشر المعتمدة في المجال، ثم نختم الباب بخلاصته فمراجعه.

٢- عمليات التعرف الآلي على الكتابة

تبدأ عمليات التعرف الآلي (والتي تشمل عملية «التعرف» التي بمعنى «التصنيف» وما يسبقها ويلحقها من عمليات مصاحبة) بعد التقاط الصور وتحديد مناطق الكتابة فيها بالمعالجة المسبقة للصور (Preprocessing) وذلك لتحسين جودة ووضوح النصوص فيها، يليها -في كثير من الأنظمة- مرحلة تقطيع صور النصوص (Segmentation) إلى صور محارفها أو أي وحدات أكبر أو أصغر تناسب التعرّف. تأتي بعد ذلك مرحلة استخلاص الملامح (Feature Extraction) التي تُستعمل لاحقا في التصنيف (Classification) بعد تدريب المصنف على ملامح أمثلة موسّمة. وأخيرا، قد تورد أنظمة التعرف الآلي مرحلة للمعالجة اللاحقة (Postprocesing) بهدف تحسين نتائج التعرف بالاستعانة باحتمالات صحتها لغويا. وكما يظهر، فبعض هذه الخطوات اختيارية قد توجد في بعض الأنظمة دون الأخرى. وفيما يلي شرحٌ للعمليات المذكورة:

١, ٢ عمليات المعالجة المسبقة

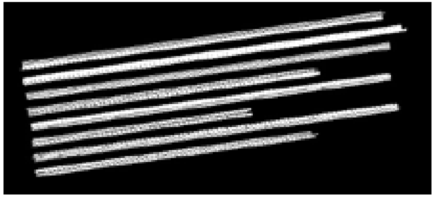
بعد تحويل المحتوى النصي إلى صورته الإلكترونية (باستخدام الماسحات الضوئية والكاميرات في حالة التحويل المتراخي أو ألواح الكتابة وشاشات اللمس في حالة التحويل الآني)، قد تُجرى بعض هذه العمليات:

- تحديد المناطق النصية في الصور (Text Localization).
- إزالة بعض التشويشات الظاهرة على الورق أو التشوهات الناتجة عن تحويل المحتوى النصي إلى إلكتروني (Noise Removal) [٨،١٢،١٣]. وللتشويش أنواع من أشهرها في مجالنا «تشويش الملح والفلفل»، وهو اسم لطيف لانقلاب بعض العناصر الصورية (Pixel) إلى اللون الأبيض أو الأسود.
- تمثيل الصورة باللونين الأبيض والأسود بدلا من تدرجات الرمادي والألوان، وهو ما يعرف باسم الترميز الثنائي، حيث يتم اعتماد قيمة من اثنتين فقط لكل

عنصر صوري (عادة ما نرمز لهما بالصفر والواحد) ليمثل أحدهما ما يظهر داكنا كالحبر ويمثل الآخر ما يظهر فاتحا كخلفية الصفحة.

وثمة العديد من تقنيات تحويل الصور إلى ثنائية، يحدد في بعضها لون العنصر (أبيض أم أسود) من قيمته هو فقط حسب حد فاصل (ثابت أو متأقلم) ويحدَّد بعضُها بموجب قيمة العنصر الصوري المراد تثنيته وقيم ما يجاوره من عناصر صورية أيضا.

- تصحيح انحراف الكتابة (سواء أحدث الانحراف عند الكتابة أو عند تحويل الصفحة إلى صورة إلكترونية) يوضحه شكل ٦. وتبدأ عملية تصحيح الانحراف عادة بتقدير درجة الانحراف، قبل أن يُدوَّر النص في الاتجاه المعاكس لانحرافه وبزاوية مساوية لزاوية الانحراف. ولتقدير زاوية الانحراف، كثيرا ما تستخدم تقنياتٌ مبنية على حساب الإسقاطات (Projections) (أي مجموع العناصر الصورية الغامقة في كل من أعمدة أو أسطر الصورة)، أو «تحليل المكونات الرئيسية» (Principal Component Analysis) أو هيكلة النصوص (Text Skeletonization)، أو تحليل الكونتورات المحيطة بالحروف والنصوص (Contours) أو تحويل هف (Hough Transformation) لتحديد القطع المستقيمة. شكل ٦ وشكل ٧ تعرض أمثلة صورية لإيضاح بعض هذه التقنيات المساندة لتصحيح انحراف الكتابة واستخراج الملامح ومعالجة الصورة.



(أ) (ب)

شكل (٦): تحويل هف (أ) قبل و(ب) بعد تطبيقه على نص لاتيني [١٧] وعربي [١٨].

	(أ)	(ب)
هيكلة النصوص		
تحليل الكونتور		

شكل (٧): كتابة بخط اليد (أ) قبل و(ب) بعد هيكلة النصوص [١٤] وتحليل الكونتور [١٥،١٦].

- ثمة عملية معالجة مسبقة أخرى تتعلق بميل أجزاء الحروف الصاعدة والنازلة عن الاتجاه الرأسي، وذلك أن بعض الحروف قد تظهر في بعض المواضع مائلة، إما لإبرازها كما يحدث عند استعمال خاصية الخط المائل (Italic) أو بسبب وضعية اليد عند الكتابة. وعادة ما يراد في هذه الحالة تعديل زوايا الأجزاء الرأسية إلى زاوية موحدة (غالبا ما تكون الزاوية العامودية) للتخفيف من الاختلافات بين أشكال الحروف في مواضعها المتعددة. تسمى هذه العملية بتعديل الميل (Slant Correction).

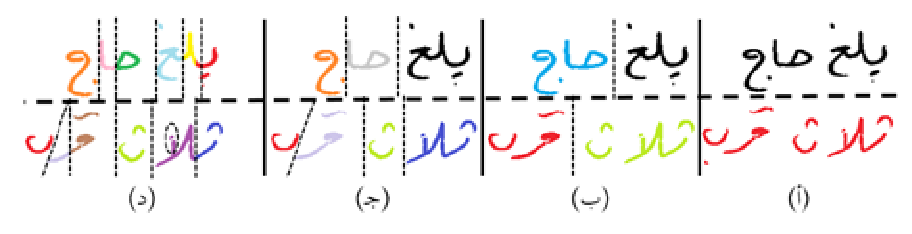


شكل (٨): كلمة «ليان» (أ) قبل تعديل الميل و(ب) بعد تعديل الميل [١٩].

- وأخيرًا يمكن السعي لتقليل التباين في أحجام الخطوط عبر ما يسمى بضبط حجم الحروف وتطبيعه (Size Normalization)، مثل سعي بعض الطرق [٢٠] لتوحيد ارتفاعات الأجزاء الصاعدة من الحروف وضبط أحجام بقية الحروف بالتناسب مع ذلك. هذا، ويجدر التأكيدُ على أن وجود -فضلا عن ترتيب- خطوات ما قبل المعالجة ليس موحدا، بل متباين من نظام لآخر.

٢,٢ التقطيع

تقطيع صور النصوص (Segmentation) عملية تهدف للحصول على صور «مقاطع» أو أجزاء أساسية من النصوص (كالحرف بالنسبة للكلمة مثلا) [٢١]. قد يجري «التقطيع» على عدة مستويات، كتقطيع صور النصوص إلى أسطر، وتقطيع الأسطر إلى كلمات أو دون-الكلمات (Subwords) («دون-الكلمات» هو ما يكتب متصلا في العربية لعدم انتهاء الكلمة وعدم ورود حرف لا يتصل بما بعده ضمنها، وتسمى أيضا Pieces of Arabic Words أو Connected Components في أحد معنييها). بل إن عملية تحديد مناطق النصوص في الصور (Text Localization) -المذكورة آنفا ضمن عمليات المعالجة المسبقة- قد ينظر إليها على أنها من مستويات تقطيع صور الصفحات لقراءتها آليا. بيد إن أشهر مستويات التقطيع وأهمها على الإطلاق والمراد بمصطلح «التقطيع» إذا أُطلق: هو تقطيع صور النصوص إلى محارفها (Character Segmentation) حيث إنه -إذا نجح- قد يسهل عمليات التعرف الآلي اللاحقة. يوضح شكل ٩ بالألوان نتائج التقطيع: إلى أسطر وكلمات ودون-الكلمات والمحارف.



شكل (٩): تقطيع صور النصوص إلى (أ) أسطر و(ب) كلمات و(جـ) دون-الكلمات و(د) محارف [٢٢][٢٣].

فصورة النص إذا كانت تحوي عدة أسطر فقد يراد تقسيمها كل سطر على حدة. وهذه الخطوة قد تزداد صعوبة للفقرات المائلة أو التي في جوانبها هوامش كما في الكثير من المخطوات الأثرية [٢٢]. لذلك، قد يتوجب استعمال أساليب أكثر ذكاء في هذه الحالات كي نتمكن من تقطيع الأسطر كلها دون دمج مكونات عدة أسطر سويا (Under-Segmentaion)، ودون تقطيع سطر ما إلى عدة أسطر (Over-Segmentation) ودون توزع مكونات السطر إلى أسطر مجاورة (Miss-Segmentation). وهذه هي الأنواع الثلاثة لأخطاء التقطيع عموما: عدم تقطيع ما حقه التقطيع، والإفراط في تقطيع ما ليس حقه التقطيع، والخطأ في موضع التقطيع).

غالبا ما تُقطع الأسطر إلى كلمات بناء على المسافات البيضاء بينها، وإن كانت الكلمات العربية قد توجد في بعضها فراغات بيضاء أصغر بين أجزائها المتصلة، مما قد يصعب تقطيعها. ثمة متعرفات تحاول قراءة الكلمات كليا (Holistic) بموجب بعض ملامحها دون اللجوء للتقطيع الحرفي الكامل لها، وهو كما يحدث عند استنتاج القارئ المتمرس للكلمات رغم خطأ ترتيب بعض حروفها، كما في المثال الذي في شكل ١٠.

I cdnuolt blveiee taht i cluod aulaclty uesdnatnrd waht i was rdanieg. The phaonmneal pweor of the hmuan mnid, aoccdrnig to a rscheearch at cmabrigde uinervtisy, it dseno't mtaetr in waht oerdr the ltteres in a wrod are, the olny iproamtnt tihng is taht the frsit and lsat ltteer be in the rghit pclae. The rset can be a taotl mses and you can sitll raed it whotuit a pboerlm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe. Azanmig huh? Yaeh and i awlyas tghuhot slpeling was ipmorantt!

شكل (١٠): نص إنجليزي مقروء رغم خلط ترتيب حروف الكلمات الداخلية.

تتجلى معضلة كمعضلة «البيضة والدجاجة» بين عمليتي تقطيع النص العربي إلى محارفه والتعرف عليه، إذ يصعب تقطيع المحارف دون تعرّف عليها، بينما يصعب التعرف على النصوص دون تقطيعها لمحارفها! لذا، لم تنجح أكثر أنظمة القراءة الآلية المعتمدة على تقطيع الحروف، وظهرت أنظمة تداخل التقطيع مع التعرف وتناوبُهما لتحاكي قراءة الإنسان، كما ظهرت أنظمة تدعو للتعرف على دون-الكلمات العربية.

يطلق مصطلح «الجزء المتصل» في سياق التعرف الآلي على الكتابة العربية بمعنيين: ما يشمل النقاط والتشكيل ضمن محارفه (وهو يُرادف «دون-الكلمات»)، وأيضا ما هو مجرد عن النقاط والهمزات والمدة والتشكيل، مع جعل النقاط والهمزات والمدة والتشكيل أجزاء متصلة مستقلة.

تتنوع أضرب التقطيع قبل التعرف الآلي في اللغة العربية إلى أنواع، أهمها: تقطيع النص إلى محارف، وتقطيع النص إلى المكونات المتصلة، وتقطيع النص إلى كلمات للتعرف عليها كليا. ويمكن لكل من هذه الأضرب التعرف على المقاطع دون النقاط أولا ثم تحديد النتائج بالنقاط، أو التعرف عليها بالنقاط منذ البداية.

٢,٣ استخراج الملامح

تلجأ كثير من الأنظمة إلى التعبير المختصر والمركز عن الصور المراد التعرف عليها بأهم ملامحها (Features) وذلك تصغيرا لحجم البيانات وتسريعا لوقت المعالجة من جانب، وتركيزا على ما يهم القارئ من المحارف وإهمالا لما لا يهم القراءة كفروقات الخطوط الفردية، من الجانب الآخر. ومع أن تصميم واختيار الملامح المناسبة فن سبيل إتقانه هو كثير من الخبرة والتجارب وشيء من التفكر والإلهام، إلا أن ثمة اتفاق على الخصائص العامة للملامح المناسبة، أهمها:

- أن تتجاهل الفروق في كتابة الحرف الواحد (Intra-Class Variability) قدر الإمكان، إذ لا بد من اختلاف بين الكتّاب في رسمهم للحرف؛ بل إن الكاتب نفسه قد يختلف رسمه للحرف من مرة لأخرى. فالملمح المناسب يقلّ تأثره بهذه الفروقات الفردية.
- أن تُظهر الفروق بين الحروف المتعددة (Inter-Class Variability) فيعكس اختلافات أشكال «السين» و«الشين» و«الحاء» -مثلا-.
- ألا تتأثر الملامح -قدر الإمكان- بحجم الكتابة ولا بقليل من الميل والالتفاف فيها (Scale and Rotation Invariant) ولا بيسير التشويش.

اقتُبست كثير من الملامح المستعملة للعربية من أعمال وأبحاث للغات أخرى. ومن أشهر هذه الملامح: كثافة العناصر الصورية [٢٧-٣١]، وأعداد مرات الانتقال من بياض لسواد والعكس[٢٩] وملامح التدرج (gradient features) [٨،٣٠]، ومقاييس التقعرات [٨،٢٧-٢٩] وترميزات اتجاهات الس (Chain-Code Directions) [٣١،٣٢] وتوصيفات فورير (Fourrier Descriptors)[٣٣] ومرشحات «جابر» (Gabor filters) [٣٤] والملامح المعتمدة على النسب المئوية لعناصر الصورة [٨] ومؤخرا قيم العناصر الصورية مباشرة للتعلم العميق [٢٥-٢٧]. كما أن للملامح مشتقات قد تستعمل أحيانا مع الملامح الرئيسية لزيادة دقة التعرف [٨،٢٦،٢٨،٢٩]. كما قد عُرّفت بعض الملامح للتراكيب العربية أصالة تحلل نقاط النصوص وصواعد ونوازل الحروف [٢٩،٣٥].

٤ , ٢ التصنيف

عملية التصنيف (ويطلق عليها مجازا «التعرف») تهدف لمعرفة رمز النص من ملامحه بعد تعلمه من أمثلة. تمر المصنفات بمرحلتين على الأقل: مرحلة التدريب والنمذجة (Training and Modelling)، ثم مرحلة التعرف والتصنيف الفعلي(Recognition and Classification). كما قد تمر بعض المصنفات بمرحلة تحقق (Validation) لتحسين تدريبها ونمذجتها، وبمرحلة اختبار (Testing) لتقرير نسب نجاحها في البحوث العلمية والمسابقات.

- **التدريب**

يعطَى المصنف في مرحلة التدريب أمثلةً مُوسّمة (Labeled) برموز المحارف أو الكلمات التي في تلك الأمثلة، وذلك حتى «يتعلم» النظام -بإحدى خوارزميات التعلم- أن يسم أمثلة لم تعرض عليه حسب ملامحها. ينتج عن مرحلة التدريب «نماذج» يستعملها المصنف لاحقا في مرحلة التعرف.

- **التعرف**

وهي المقصود النهائي للقارئات الآلية، والوحيدة التي تهم المستخدم النهائي. يُعطى المصنف في مرحلة التعرف الملامح المراد التعرف على نصوصها، وهذه هي المرحلة الوحيدة التي لا تستعمل فيها أوسام مسبقة للنصوص.

- **الاختبار**

تأتي مرحلة الاختبار -بعد أن يجهز المتعرف- لقياس مدى نجاحه، فيُعطى صور المحارف دون أوسمتها، ويُحتفظ بالأوسمة للمقارنة بها وتقرير نسب النجاح (التعرف الصحيح) والخطأ. تفصّل نسب الخطأ أحيانا إلى أخطاء إدراج (Insertion Errors) وأخطاء إسقاط (Deletion Errors) وأخطاء تبديل (Substitution Errors). تُجرى مرحلة الاختبار عادة على قواعد بيانات مشهورة لتَتَيَسَّر المقارنة بين البحوث.

قد يعيد البعض استخدام جزء من صور التدريب في الاختبار، بينما يحبذ آخرون الفصل التام بين أمثلة التدريب وأمثلة الاختبار للتقليل من احتمالية «الحفظ الجامد» دون تعلم (Overfitting). وقد ترجح هذه الطريقة أو تلك حسب حجم البيانات المتوفرة، وحسب الهدف من التعرف (هل هو محدود بخطوط كتّاب معيّنين أو عام).

وبينما لا بد أن تشمل صور التدريب الموسومة جميع أنواع المحارف وأشكالها، لا يشترط ذلك لصور الاختبار (وإن كان قد يفضل). ويختلف الباحثون في نسب ما يخصصون من البيانات للتدريب والاختبار، وينصح أن تكون تلك النسب قريبة من ٦٠٪ للتدريب و٤٠٪ للاختبار [٣].

- **التحقق**

نستطيع توضيح مفهوم التحقق بموجب مرحلة الاختبار: فالتحقق ما هو إلا «اختبار تجريبي» يهدف لتلافي مواضع الضعف وتحسين أداء المصنف بناء على نتائج مؤقتة لا يُهدف لنشرها. يساعد التحقق الصحيح في تجنب بعض المحاذير مثل «الحفظ الجامد» (حيث يُفرِط المصنف في «قولبة» الفروقات والتشابهات التي مثلتها له ملامح أمثلة التدريب) فينكشف ذلك عندما تعطى له أمثلة التحقق، مما يسمح بتدارك الأمر وإعادة النمذجة. وخلافا لمرحلة الاختبار، فإن مرحلة التحقق يمكن أن تكرر مرارا.

٥,٢ المعالجة اللاحقة

قد يستعان في الخطوات الأخيرة للتعرف الآلي بمعاجم (Lexicons) وقواعد لغوية (Linguistic Rules) لما تقبله اللغة أو ترفضه، أو بنماذج إحصائية (Statistical Models) للشائع لغويا كـ«الورودات الأقرب» (N-Grams)، لترجيح أو استبعاد نتائج التعرف، لا سيما عندما تكون الكلمات المراد التعرف عليها محصورة في مجال محدد كالطب أو الهندسة أو أسماء مدن (Domain-Specific).

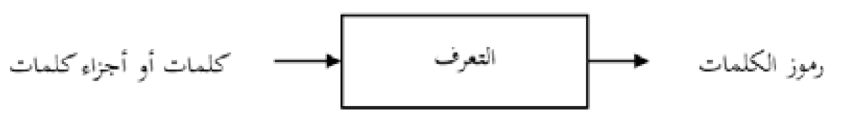
٣ - التعرف على الكتابة حسب علاقة التصنيف بالتقطيع

ألمحنا -سابقا- إلى حدوث «الدّور» (Recursion) بين التقطيع والتعرف، وقد نشأت عن هذه المعضلة أنواع لمعماريات القراءة الآلية، منها:

١,٣ التعرف القائم على التقطيع

التعرف القائم على التقطيع هو الأسلوب التقليدي حيث تُقطع صور النصوص إلى صور للوحدات التي يراد التعرف عليها قبل عملية التصنيف [٣٦]. ويُعرَف هذا الأسلوب أيضا بأسلوب التقطيع الخارجي (External Segmentation)، أو التقطيع الصريح (Explicit Segmentation)، وشكل ١١ يوضح معماريته العامة.

وأكثر ما قد تنجع فيه هذه الطريقة للمطبوع من النصوص، لا سيما إذا كان بخطوط صممت خصيصا لذلك، كالخطوط التي تعطي جميع المحارف نفس العرض أو التي تترك فراغا صغيرا بين المحارف [٢٤].



شكل (١١): مخطط عمليات التعرف القائم على التقطيع.

٢,٣ التعرف الكلي (دون التقطيع إلى محارف)

اقترح باحثون [٣٧] التعرف على الكلمات أو أجزاء الكلمات العربية دون تقطيعها إلى محارف، كما في المعمارية المبينة في شكل ١٢.

كلمات أو أجزاء كلمات ← التعرف ← رموز

شكل (١٢): مخطط عمليات التعرف الكلي.

ومما يساند هذا الاتجاه: تغيّب الحركات عن أكثر كتاباتنا اليومية (حيث يتدرب القارئ العربي على استنتاج التشكيل والكلمات من السياق) ولمن لا يعرف العربية، قام بعض الباحثين [٣٨] بتقديم مثال إنجليزي حذفوا منه جميع حروف العلة (Vowels) لتوفير تجربة شبيهة بقراءة العربية، وهو ما أوردناه للفائدة في شكل ١٣. يؤدي تغيّب الحركات في كتاباتنا اليومية إلى «إعادة تدوير» رسم الكلمات، فمثلا رسم «كتب» يستعمل لكلمات عديدة مثل «كَتَبَ» و«كُتِبَ» و«كُتُبٌ» و«كُتُبٍ»، والتي لو كانت بالتشكيل أو بأحرف لاتينية لاحتاجت لأربعة أصناف (".kataba". "kutiba" .("kutubun". "kutubin").

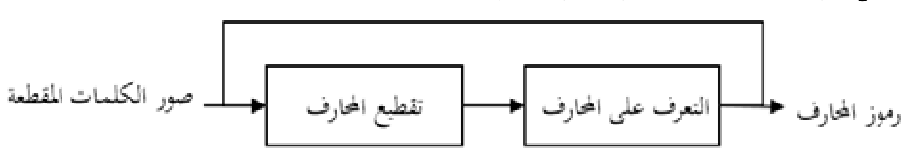
علاوة على ذلك، يتجه الكثير من الباحثين لحذف النقط والهمزات والمدة من صور النصوص ليشمل الرسم الواحد أكثر من كلمة، فتدخل تحت صنف «كتب» عندئذ كلمات مثل (كثْب، كنب، كبت، كئب).

Just to feel the task, read the following English sentence:
"jst t fl th tsk, rd th fllwng nglsh sntnc"

شكل (١٣): مثال إنجليزي حذفت منه حروف العلة [٣٨].

٣,٣ التعرف الذي يتخلله تقطيع ضمني

التقطيع القائم على التعرف، أو التقطيع الداخلي (Internal Segmentation) أو الضمني (Implicit Segmentation)، يستند إلى خوارزميات تقترح أثناء التعرف مواضع أولية لابتداء وانتهاء المحارف، ثم تكرر محاولات التعرف إلى الحصول على نتائج جيدة إحصائيا أو لغويا. شكل ١٤ يوضح معمارية التقطيع القائم على التعرف. ويمكن التجوز واعتبار أن التعرف الضمني يجعل التقطيع والتعرف يحدثان معا في نفس الوقت، كأشبه ما يكون بقراءة الإنسان.



شكل (١٤): مخطط عمليات التعرف الذي يتخلله تقطيع ضمني.

وقد أخرنا الكلام عن هذا النوع لأهميته حتى نتمكن من الاستفاضة بطريقتي «نماذج ماركوف الخفية» وتقنيات «التعلم العميق» العاملتين بالتقطيع الضمني.

١,٣,٣ التعرف بنماذج ماركوف الخفية

نماذج ماركوف الخفية (Hidden Markov Models أو HMM ختصارا) تعمل عادة على صور الأسطر الكاملة، رغم وجود القليل من الأعمال التي استعملت نماذج ماركوف الخفية في التعرف على أعداد ومحارف منفصلة أو مقطعة [٣٩،٤٠] أيضا.

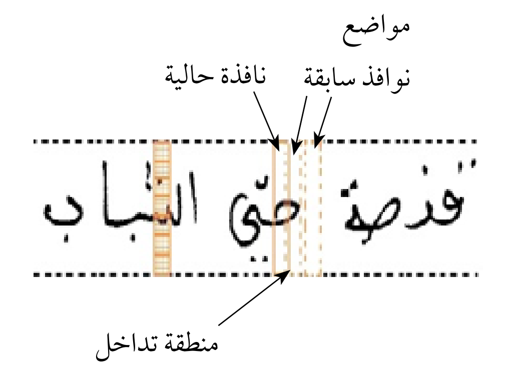
يرجح كفة استعمال HMM على أساليب أخرى (كآلات متجهات الدعم (Support Vector Machines) والشبكات العصبية الاصطناعية (Artificial Neural Networks) والغابات العشوائية (Random Forests)) هو قدرتها على تقطيع صور النصوص ضمنيا أثناء التعرف على صور الأسطر وأحيانا الفقرات.

نتحدث فيما يلي عن أشهر طرق استخراج الملامح المستعملة مع نماذج ماركوف الخفية، ثم نتطرق لوحدات النمذجة المشتهرة فيها ولشكل السلاسل الأكثر استعمالا (وهو ما يسمى بـ«طوبولوجيا» (Topology) السلسلة) ثم نذكر أشهر خوارزمياتها للنمذجة الصورية (للتدريب) واللغوية (للمعالجة اللاحقة).

الملامح الأشهر استعمالا مع متعرفات نماذج ماركوف الخفية

عادة ما تلجأ أنظمة التعرف القائمة على نماذج ماركوف الخفية -بعد عمليات المعالجة المسبقة- إلى حساب الملامح عبر ما يعرف «بـالنافذة المنزلقة» (Sliding Window) [٤١، ٣٠-٤٣]؛ حيث يحدد جزء له نفس ارتفاع صورة السطر المراد التعرف على محتواه النصي بعرض مقارب لذلك الارتفاع، فتحسب الملامح ذلك الجزء من الصورة والذي يعرف باسم «النافذة». تُزلق النافذة (تزاح) من أول السطر (يمينه) حتى آخره (يساره) وتكرر عملية حساب الملامح مع كل موضع من مواضع النافذة.

ثمة أسلوبان مشهوران لإزاحة النوافذ المنزلقة، أحدهما: إزاحتها بمقدار عرض النافذة بحيث لا يحصل تداخل بين مواضع النوافذ [٢٧]، والآخر: إزاحتها بعرض أقل من ذلك فيحصل تداخل جزئي بين النوافذ [٤٤، ٤٣، ٣٠] كما هو مبين في شكل ١٥ [١٨].



شكل (١٥): النافذة المنزلقة ويرى فيها تداخل بين النافذة الحالية (المستطيل الأخير) وبعض السابقة (المستطيلات المنقطة) [١٨].

عادة ما تكون النوافذ المنزلقة ذات عرض ثابت، بيد إن بعض التجارب أُجريت لجعل عرض النافذة يتغير اعتمادا على خصائص الصورة -كأبعاد القيعان التي بينها قمم في إسقاطات عناصرها الصورية الرأسية (Vertical Projection)- وقد وجد هؤلاء الباحثون تحسنا في النتائج جراء ذلك [١٢]. كما أن ثمة تجارب استحدثت نوافذ منزلقة مائلة (يمينا ويسارا) استعملت مع نماذج ماركوف الخفية [٤٥، ٢٩]. وأخيرا، فعلينا أن ننوه إلى عدة أبحاث سعت لاستعمال نماذج ماركوف الخفية مجردة عن أسلوب النوافذ المنزلقة بالكلية [٤٧، ٤٦].

وحدات التعرف الأشهر في نماذج ماركوف الخفية

أكثر نماذج ماركوف الخفية تستهدف المحارف [١٢،٢٩،٣٠،٤١،٤٨] لا الحروف [٤٦] وهذا -كما أسلفنا- لأن الحرف العربي له أكثر من شكل بحسب اتصاله بما قبله وما بعده مما يصعب نمذجتها كلها بسلسلة موحدة؛ فأكثر المحارف استعمالا أشكال الحرف حسب الموضع («الشكل المنفصل»، و«الشكل الابتدائي»، و«الشكل المتوسط»، و«الشكل النهائي»)، ولكن الأمر لم يخلُ من مساع لتجريب محارف أخرى كنمذجة الأجزاء الرئيسة والمكملة من الحروف إلى محارف [٤٧]، ونمذجة الأجزاء التي تشترك فيها عدة محارف [٤٩،٥٠]، ونمذجة حرفين أو عدة حروف معا (كما في سعينا لنمذجة المتراكبات الأشهر لمحرفين فأكثر معا [٥١]). وقد وجدت أيضا مساعٍ لنمذجة المحارف وأجزاء الكلمات مجردة من النقط والهمزات والتشكيل، ونمذجة تلك النقاط والهمزات والتشكيل على حدة [٥٢]، وأخرى لنمذجة الفراغات البيضاء (بين الكلمات) أو البينية (التي تقع بين أجزاء الكلمة المنفصلة) [٢٥،٢٩] بسلاسل مستقلة.

أشكال السلاسل الأشهر لنماذج ماركوف الخفية

تعتمد أغلب نماذج المحارف على طوبولوجيا باكيس (Bakis) لسلاسل ماركوف الخفية، وهو اسم للطوبولوجيا التي تسمح دائما بتخطي المرحلة التالية من السلسلة إلى التي بعدها كما يبينها شكل ١٦ (ب). والسر في اللجوء لهذه السلاسل يكمن في مرونتها، خاصة وأن حروف العربية تتفاوت في العرض، بل إن عرض الحرف نفسه قد يختلف من موضع لآخر بسبب استعمال التطويلات أو أسلوب خط معين. وهذا لا يعني عدم وجود بحوث ناجحة استعملت طوبولوجيات أبسط (كالخطية [٣٠،٤٢،٤٤]) وأعقد [٣٢] من طوبولوجيا باكيس، لكن المقصود الإشارة لما اتفق على استعماله أكثر الباحثين.

(ب) (أ)

شكل(١٦): (أ) طوبولوجيا خطية (ب) طوبولوجيا باكيس.

الخوارزميات الأشهر لتدريب نماذج ماركوف الخفية

بعد نمذجة المحارف، يكثر استعمال خوارزمية فيتيربي (Viterbi)، والتي ظهرت لأول مرة في منتصف الستينيات من القرن الماضي، لتحديد التسلسل الأمثل لنماذج ماركوف اعتمادا على البرمجة الديناميكية (Dynamic Programming) [١٢،٤١،٤،٥٣،٥٤].

ولتحديد احتمالات الانتقال الأنسب بين النماذج، فكثيرا ما تُستعمل خوارزمية تدريب تعرف هي الأخرى باسم مخترعيها، باوم وويلتش (Baum-Weltch) [٥٥،٤٦،٤٣،٣٢،٣٠،٨]. وللمزيد، يمكن الرجوع لبحث « تقنيات التعرف الآلي على الكلام المنطوق وتطبيقاتها في القرآن الكريم-واقع وطموح» في كتابنا السابق «الحرف العربي والتقنية» [٥٦].

النمذجة اللغوية

قد يستعان بعد التصنيف بنوع آخر من النماذج لترجيح كفة المقبول والشائع احتماليا ولغويا من الكلمات، وهذه تسمى بالنماذج اللغوية. ونماذج ماركوف الخفية تتيح استعمال النماذج اللغوية (وخاصة «الورودات السابقة» n-grams) بسلاسة. فقد استعملت هذه النماذج على مستوى الحرف والمحرف إلى ٤ ورودات سابقة بعدما حسبت من «مدونات لغوية» حوت عشرات الملايين من الكلمات [٨، ٥٧]. كما قد تستعمل أيضا أجزاء-الكلمات للنمذجة اللغوية [٥٨] ضمن تنويعات أخرى لوحدات النمذجة اللغوية.

٢,٣,٣ التعلم العميق للتقطيع ضمني

موجة «التعلم العميق» (Deep Learning) صاعدة -هذه الأيام- في شتى مجالات رؤية الحاسب (Computer Vision)، وليس التعرف على النصوص مستثنى من ذلك [٥٩،٦٠]. التعلم العميق قائم على الشبكات العصبية الاصطناعية (Artificial Neural Networks أو ANN) كثيرة الطبقات. وإذا استعين فيها بتقنيات للتصنيف الزمني التوصيلي (Connectionist Temporal Classification أو CTC) والشبكات العصبية المتكررة (Recurrent Neural Network أو RNN) وبُنى الذاكرة قصيرة-

المدى الطويلة (Long Short-Term Memory أو LSTM)، تصبح ماهرة في التقاط السياقات المشتركة -وإن تباعد ورودها-، وهو ما يجعلها مفيدة في مجال التعرف على النصوص دون التقطيع المسبق [٥٩].

لذا، فقد فتحت هذه التقنيات الحديثة الباب للتعرف على النصوص العربية المكتوبة بالتقطيع الضمني كما في [٢٦،٥٩،٦٤]. فالشبكات العصبية التكرارية للتعرف الآلي على النصوص دون تقطيع يمكن أن تستخدم على النهج الذي قدمه Graves و Schmidhuber ثم عاد فطوره Graves لاحقا، بحيث تعالج العناصر الصورية في شبكات عصبية تكرارية متعددة الأبعاد (Multi-Dimensional Recurrent Neural Network أو MDRNN اختصارا).

٣,٣,٣ نماذج ماركوف الخفية مع التعلم العميق

تنتشر طرق هجينة لاستعمال التعلم العميق جنبا إلى جنب مع نماذج ماركوف الخفية (كما في [٥٨] و[٦٢]). فنماذج ماركوف الخفية قد تستعمل في محاذاة المحارف مع صورها (التقطيع الضمني) -مثلا- قبل التعلم العميق للمحارف، أو لتدريب الشبكات العصبية التكرارية مع البنى ثنائية للذاكرة قصيرة-المدى الطويلة على نتاج تلك المحاذاة القسرية، قبل استعمال نتائج هذه الأخرى لإعادة تدريب نماذج التعرف النهائي على النصوص [٢٦][٦٤][٤٢].

وقد قُدّمت دراسات قارنت بين نظم مبنية على التعلم العميق (كالشبكات العصبية التكرارية مع الملامح ذات الذاكرة قصيرة-المدى الطويلة (أي RNN مع LSTM)) للقراءة الآلية العربية وأخرى على نماذج ماركوف الخفية [٦٥] باستعمال نفس الملامح، استطاع في أغلبها نظام التعلم العميق التفوق على غيره. وخلصت دراسة مقارنة إلى أنه ليس ثمة فرق كبير بين استعمال الشبكات العصبية التكرارية مع الملامح ذات الذاكرة قصيرة-المدى الطويلة وبين استعمال الشبكات العصبية العادية (MLP)، وأنه لا فرق يذكر كذلك بين استحداث واستعمال ملامح متقدمة وبين تقديم العناصر الصورية بصورتها الخام إذا تم تدريب كل نظام بما يناسبه [٦٦]. وهذا مما يحث الباحثين إلى محاولة تفهم هذه الأنظمة وخصائص كل منها بشكل أكبر، ويدعوهم للتدقيق في نتائج التعرف والمقارنات بينها، وهو ما لا يتأتى إلا بفهم طبيعة البيانات التي تتم

عليها مرحلة الاختبار. لذا، فقد أفردنا الفصل التالي للتعريف بأشهر قواعد بيانات الكتابة العربية اليدوية التي تستعمل في تدريب واختبار وتقرير نتائج المتعرفات الآلية في البحوث العلمية.

٤- مقارنات لبعض أشهر أنظمة التعرف الآلي على النصوص العربية المكتوبة بخط اليد

قبل المقارنة بين النتائج المنشورة لأي متعرفات، يجدر بنا التعرف على قواعد البيانات التي تُجرى اختبارات كفاءة المتعرفات الآلية عليها. وفيما يلي، نذكر أشهر هذه القواعد مع نُبَذٍ عنها، ثم بعد ذلك، نقدم جداول لأشهر أنظمة التعرف الآلي على النصوص المكتوبة بخط اليد، مع إيراد نسب الخطأ فيها، والإشارة إلى خصائصها وقواعد البيانات التي قُررت نتائجها عليها.

٤, ١ قواعد بيانات للكتابة العربية اليدوية

نعرض فيما يلي تسع قواعد بيانات -مرتبة حسب وقت نشرها التقريبي- مع نبذة عن كل منها. ثم نعرض بعدها مقارنة جدولية لها.

٤, ١, ١ قاعدة بيانات الإسراء

تتكون قاعدة بيانات الإسراء (AL ISRA database) [٦٧] من كلمات عربية وأرقام وتوقيعات وجمل حرة، جُمعت من حوالي مائة طالب من جامعة الإسراء الأردنية لصالح باحثين في جامعة كولومبيا البريطانية. ولكن -وكأغلب قواعد البيانات حينها- فقد افتقرت قاعدة الإسراء إلى فقرات نصية كاملة مكتوبة في بيئة طبيعية.

٤, ١, ٢ قاعدة بيانات الشيكات العربية

جمع العوهلي وآخرون قاعدة بيانات لسندات مصرفية (شيكات) عربية (CENPARMI) [٧] والتي اشتملت على نصوص وأرقام تم استخراجها من ٣٠٠٠ سندا وفرها مصرف الراجحي في المملكة العربية السعودية. لذا، فهذه القاعدة قد تفيد كثيرا تطبيقات التعرف على محتوى السندات البنكية.

٣, ١, ٤ قاعدة بيانات النص العربي المكتوب بخط اليد

تتضمن قاعدة بيانات النص العربي المكتوب بخط اليد (Database for Arabic Handwritten text أو AHDB) الكلمات المستخدمة في كتابة المبالغ على السندات المصرفية كما اشتملت أيضا بعض صفحات الكتابة الحرة بخط ١٠٠ كاتب [٦٨].

٤, ١, ٤ قاعدة البيانات الحرة للأحرف العربية بخط اليد

أعد الباحثان خضر وعبندة [٤] قاعدة بيانات مبكرة للأحرف العربية كتبت بشكل حر (بلا قيود على طريقة الكتابة) من قبل ٤٨ كاتبا. كما طلبوا كتابة فقرة نصية حوت أشكال المحارف والأعداد العربية وبعض الرموز دون فرض قيود على طريقة الكتابة.

٥, ١, ٤ قاعدة بيانات الأرقام، والحروف، والرموز المعزولة والمتصلة في كلمات

وهي قاعدة بيانات طورها العمري لتحوي أرقاما، وحروفا، ورموزا كتلك التي تكتب في التواريخ والأعداد والكلمات [٦٩].

٦, ١, ٤ قاعدة بيانات الأعداد العربية

قاعدة بيانات الأعداد العربية المكتوبة بخط اليد (A database of handwritten Arabic أو ADBase) مناسبة لأهداف التعرف الآلي على الأرقام العربية (والتي تعرف أيضا بالأعداد الهندية) [٧٠].

٧, ١, ٤ قواعد بيانات مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا

مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا (Multilingual Automatic Document Classification. Analysis and Translation أو MADCAT) ترعاه الوكالة الأمريكية لمشاريع الأبحاث المتقدمة (DARPA) بهدف دعم الجيش الأمريكي بقدرات على القراءة والترجمة الآلية من عدة لغات، من أهمها العربية [٧١]. أنتج المشروع في مراحله الأولى نصوصَ تدريبٍ مكتوبة بخط اليد، حيث تعاونت الوكالة مع اتحاد البيانات اللغوية (Linguistic Data Consortium أو LDC) لإنشاء قاعدة البيانات التي حوت ٩٦٩٣ صفحة مكتوبة بخط اليد، شملت وثائق عربية أصلية (رمّزت وقطّعت إلى أسطر، ومسحت ضوئيًا بدقة ٦٠٠ نقطة في البوصة، ووسمت أجزاؤها، وترجمت نصوصها إلى الإنجليزية).

مصادر الوثائق حوت -في الغالب- من الأخبار والمدونات الإلكترونية. وقد أقيمت مؤخرا مسابقات (NIST-OpenHaRT) [٧٢] للتعرف على أسطر نصية، وقد أتيحت جزئيا لبعض الباحثين، ولكن ما تزال أكثر توزيعاتها وأساليب مقارنة النتائج سرية فيها [٧٣،٧٢،٦٤].

٨, ١, ٤ قاعدة بيانات IFN/ENIT

تعاون كل من معهد تكنولوجيا الاتصالات (Institut für Nachrichtentechnik أو IFN) في جامعة براونشفايغ التقنية (Technische Universität Braunschweig) في ألمانيا مع المدرسة الوطنية الهندسية في تونس (Ecole Nationale d'Ingénieurs de Tunis أو ENIT) لإخراج ما كان حينا من الدهر المرجعيةَ الأكثر شيوعًا لتقرير نتائج بحوث التعرف على النصوص العربية المكتوبة بخط اليد [٧٤]؛ وقد يرجع السبب في انتشارها لتوفرها مجانا منذ نشأتها ولنشاط منشئيها في خدمتها وعقد المسابقات عليها، فقد تم استعمال قاعدة IFN /ENIT في مسابقات عديدة للتعرف الآلي على النصوص العربية، عرضت نتائجها في مؤتمرات مهمة [٧٥-٧٩].

تتكون هذه القاعدة من صور مكتوبة بخط اليد لأسماء ٩٣٧ مدينة وبلدة تونسية (أي أنها تضمنت معلومات للخدمة البريدية في الأصل) مقسمة إلى سبع مجموعات (A. B. C. D. E. F. S بعد إضافة المجموعتين F وS مؤخرا). وتعتبر مجموعة S الأكثر صعوبة لأنها لم تجمع في نفس بيئة بقية المجموعات، فتضمنت أنماطا كتابية مختلفة عن المجموعات الأخرى.

٩, ١, ٤ قاعدة بيانات «خط»

تُعنى قاعدة بيانات «خط» (KFUPM Handwritten Arabic TexT database أو KHATT) بالكتابة الحرة [٨٠،٦]، حيث تحوي فقرات كتبها ١٠٠٠ شخص (كلّ منهم كتب أربع فقرات، اثنتان منها نصهما موحد).

تحوي القاعدة صور الفقرات مقطعة على مستوى أسطر النصوص، وتنقسم إلى ثلاث مجموعات: مجموعة التدريب (٤٨٠٨ سطرا)، ومجموعة التطوير (٩٣٧ سطرا)، ومجموعة الاختبار (٩٦٦ سطرا).

١٠,١,٤ جدول قواعد بيانات الكتابة العربية اليدوية

نلخص في الجدول أدناه وصف وعدد كتاب بعض قواعد البيانات المذكورة آنفا.

جدول (١): ملخص بعض قواعد بيانات الكتابة العربية.

مختصر اسم قاعدة البيانات مع إحالة مرجعية	وصف مختصر	عدد الكتاب
IFN/ENIT [٧٤]	٢٦,٤٥٩ صورة لأسماء مدن وبلدات تونسية	٤١١
الإسراء Al-Isra [٦٧]	٣٧,٠٠٠ صورة كلمة و ١٠,٠٠٠ صورة عدد و٢٥٠٠ صورة توقيع و٥٠٠ صور لجمل	٥٠٠
MADCAT [٤٢]	٩,٦٩٣ صور لصفحات من وثائق أخبار وغيرها	٤٠٠ على الأقل
شيكات مصرف الراجحي CENPARMI [٧]	٣,٠٠٠ صورة لقيم شيكات بالأرقام والحروف	-
AHDB [٦٨]	١٠,٠٠٠ صورة لمصطلحات شيكات مصرفية	١٠٠
Khedher et al. [٤]	صور حروف	٤٨
العمري [٦٩] Alamri et al.	٤٦,٨٠٠ صورة لأعداد و١٣,٤٣٩ صورةلأرقام (سلاسل عددية) و ٢١,٤٢٦ صورة لحروف و١١,٣٧٥صورة لكلمات و١,٦٤٠ صورة لرموز كتابية خاصة وعلامات ترقيم	٣٢٨
الأرقام العربية ADBase [٧٠]	٧٠٠,٠٠٠ صورة لأعداد	٧٠٠
قاعدة بيانات «خط» KHATT [٦،٨٠]	١,٠٠٠ صورة نموذج و٢,٠٠٠ صورة لفقرات كتابة حرة	١٠٠٠

وبعد تعرفنا على بيانات الاختبار نستطيع تقديم مقارنات لأنظمة التعرف الآلي على الكتابة اليدوية العربية الأبرز في البحوث العلمية، وتقارير نتائجها حسب قواعد البيانات التي اختُبِرت عليها.

٢, ٤ مقارنة أهم بحوث المجال

نلخص هنا أهم البحوث المنشورة في مجال التعرف الآلي على النص العربي المكتوب بخط اليد. وسنقسم مناقشتنا إلى ثلاث مجموعات: الأولى لأهم بحوث التعرف على الأعداد والمحارف المنعزلة، والتعرف الكلي على أجزاء الكلمات العربية، ومحاولات مبكرة للتعرف على الكلمات كليا أو التعرف القائم على التقطيع الصريح. والمجموعة الثانية لأهم البحوث التي اعتمدت في تدريبها وتقييمها على قاعدة IFN /ENIT [٧٤] وذلك لما تتسم به هذه القاعدة من أهمية وشعبية من جانب، ولأنها محدودة الكلمات، مما يمكن من التعامل معها بأسلوب خاص. والمجموعة الثالثة لأهم بحوث التعرف الآلي على الصور النصية التي تحوي عدة كلمات ذات الخصائص اللغوية المفتوحة (ليست كلمات محدودة كما في قاعدة IFN /ENIT).

١, ٢, ٤ أهم بحوث التعرف على الأعداد، والمحارف، والكلمات، وأجزاء الكلمات المنعزلة

نقدم في جدول ٢ ملخصًا للأعمال المتعلقة بالتعرف على الأعداد والمحارف المقطعة والكلمات وأجزاء الكلمات العربية. يتيح الجدول المقارنة بين أداء أعمال ممثِّلة في هذا المجال، حيث ترتبط الجوانب الرئيسية لفاعلية التعرف بالمعالجة المسبقة، والملامح والمصنفات.

يعد التعرف على الأعداد المكتوبة بخط اليد أحد أسهل مهام التعرف إذ أن الأصناف فيها (من ٠ إلى ٩) قليلة. لذلك، نجد تقارير عن نسب نحاج بمعدل ٩٩٪ [٤٠]، أي أنها «مشكلة محلولة». أما التعرف على الأرقام (أي السلاسل التي تحوي عدة أعداد) حيث قد تتلامس الأعداد المتجاورة فما زالت اشكل تحديا وتحتاج مزيد حل [٨١]. وأهم تطبيقات التعرف على الأعداد والأرقام هي قراءة السندات المصرفية آليا.

يشبه التعرف على الأعداد سهولةً التعرف على المحارف المنعزلة؛ حيث تكمن أهم التحديات في التعرف على المحارف التي تتشابه أو تشترك في الشكل وتختلف في النقط. وأيضا يمكن اعتبار أن مشكلة التعرف على المحارف المنعزلة محلولة -إلى حد كبير- غير أن استخدامات الحروف المعزولة عمليا محدود، ربما كانت أهم تطبيقاته هي القراءة الآلية للرموز البريدية في البلدان التي تعتمد الحروف العربية المنعزلة لهذه الرموز.

وأما التعرف على الكلمات العربية -كليا أو بشيء من التقطيع- فلا تكاد تنجح إلا عندما يكون مجال المفردات الكلي لهذه الكلمات محدودًا، كما في مهام التعرف على أسماء مدن أو قيم مكتوبة خطيا.

نعرض في جدول ٢ بعض المساعي للتعرف على على الأعداد، والمحارف، والكلمات، وأجزاء الكلمات العربية المتصلة دون تقطيع يذكر.

جدول (٢): بحوث في التعرف على الأعداد، والمحارف، والكلمات، وأجزاء الكلمات العربية المتصلة.

مرجع النظام	هدف النظام	نتائج التعرف	قاعدة البيانات	سمات النظام المختصرة
Alamri et al. ٢٠٠٩ [٨١]	التعرف على الأعداد والأرقام ذات الأعداد المتلامسة	• نسبة الخطأ في التصنيف ٪١,٥٢ للأعداد غير المتلامسة • نسبة الخطأ في التصنيف ٪٧,٧٨ للأعداد المتلامسة	• صور أعداد من قاعدة CENPARMI للشيكات العربية - ٢٤,٧٨٤ صورة للتدريب - ٦,١٩٩ صورة للتقييم - ١٣٢ صورة لأرقام فيها أزواج متلامسة من الأعداد للتقييم	• نظامSVM نواته Radial Basis Function • ملامح تدرجية • خوارزمية قواعدية لفصل الأعداد المتلامسة
Awaidah and Mahmoud ٢٠٠٩[٤٠]	التعرف على الأعداد (منفصلة)	• نسبة الخطأ في التصنيف ٪٠,٨٧	• قاعدة من ٢١،١٢٠ صورة بيد ٤٤ كاتبا - ١٥،٨٤٠ صورة للتدريب - و٥,٢٨٠ صورة للتقييم	• نماذج ماركوف الخفية المنفصلة • ملامح التدرج والتقعر والميزات الهيكلية (GSC) • تقسم الصورة إلى إطارات لكل منها نفس عدد العناصر الصورية تقريبًا

مرجع النظام	هدف النظام	نتائج التعرف	قاعدة البيانات	سمات النظام المختصرة
Mahmoud and Al-Khateeb ٢٠١٠ [٨٢]	التعرف على الأعداد (منفصلة)	• نسبة الخطأ في التصنيف ١,٠٥٪ عند استخدام مصنف SVM • نسبة الخطأ في التصنيف ٢,٧٩٪ عند استخدم نماذج ماركوف الخفية • نسبة الخطأ في التصنيف ١,٢٥٪ عند استخدام الجار\ الجيران الأقرب k-NN	• صور أعداد من قاعدة CENPARMI للشيكات العربية: - ٧,٣٩٠ صورة للتدريب - ٣,٠٣٥ صورة للتقييم	• تم استعمال ٣ أنواع من المصنفات: - نماذج ماركوف الخفية HMMs - آلات متجهات الدعم SVM - الجيران الأقرب k-NN • تم استخدام ملامح مأخوذة من مرشح «جابر» اللوغاريمي (Log Gabor) بمقاييس وتوجهات المختلفة

مرجع النظام	هدف النظام	نتائج التعرف	قاعدة البيانات	سمات النظام المختصرة
Cheriet et al. ٢٠٠٧ [٨٣]	التعرف على الكلمات\ أجزاء الكلمات العربية كليا	• نسبة الخطأ في التعرف على «أجزاء الكلمات العربية» كانت ٤٧,٢٦٪	• صور أجزاء كلمات معزولة من قاعدة CENPARMI للشيكات العربية: – ٦٧ نوع من أجزاء الكلمات	• نماذج ماركوف الخفية المنفصلة – على مستوى «أجزاء الكلمات العربية» (PAWs) – أعداد مراحل السلاسل تعتمد على أعداد حروف «أجزاء الكلمات»
Dehghan et al. ٢٠٠١ [٣٢]	التعرف على الكلمات كليا	• نسبة الخطأ للكلمات: ٩٥,٣٤٪	• ١٧,٠٠٠ كلمة مكتوبة بخط اليد لأسماء ١٩٨ مدينة مختلفة، قسمت كالتالي – ٦٠٪ للتدريب – ٤٠٪ للتقييم	• نماذج ماركوف الخفية المنفصلة – سلسلة لكل كلمة – أعداد مراحل السلاسل تعتمد على معدل عرض صورة الكلمة

مرجع النظام	هدف النظام	نتائج التعرف	قاعدة البيانات	سمات النظام المختصرة
Alma'deed et al. ٢٠٠٢ [٨٤]	التعرف على الكلمات كليا	• نسبة الخطأ التقريبية للكلمات: ٤٠٪، تم استدراكها إلى ٣١٪ بالمعالجة اللاحقة	• ٤,٧٠٠ كلمة مكتوبة بخط ١٠٠ كاتب – عدد الكلمات المختلفة ٤٧ – ثلثان للتدريب وثلث للتقييم	• نماذج ماركوف الخفية المنفصلة – سلسلة لكل كلمة – أوائل المرحل تصنف الكلمات إلى مجموعات أولية
Farah et al. ٢٠٠٦ [٨٥]	التعرف على الكلمات كليا	• نسبة الخطأ التقريبية للكلمات: ٦٪	• ٤,٨٠٠ كلمة مكتوبة بخط ١٠٠ كاتب – عدد الكلمات المختلفة ٤٨ – ١,٢٠٠ للتدريب – ٣,٦٠٠ للتقييم	• عدة نظم تصنيف تجمع نتائجها لإصدار الحكم النهائي: – شبكات عصبية اصطناعية، – الجيران الأقرب – الجيران الأقرب الضبابية • الملامح هيكلية: كصواعد الكلمات ونوازلها وحلقاتها المغلقة

٢,٢,٤ أهم بحوث التعرف على قاعدة بياناتIFN /ENIT

يقدم جدول ٣ بيانات لأهم البحوث التي اعتمدت قاعدة بيانات IFN /ENIT. ونلاحظ أن استخدام المصنفات المستندة إلى نماذج ماركوف الخفية HMM هي النهج السائد لهذه الفئة. بجانب التحديات التقليدية للمعالجة المسبقة، وتطوير الملامح، والاستخدام الفعال للمصنفات؛ يجب أن يقرر المتعامل مع كلمات قاعدة بيانات IFN / ENIT وحدات النمذجة التي سيعمل عليها (الأحرف أو المحارف أو أجزاء المحارف، أو الكلمات، أو أجزاء الكلمات).

جدول (٣): بعض أهم بحوث التعرف على كلمات قاعدة بيانات IFN /ENI.

مرجع النظام	أجزاء القاعدة المستعملة للتدريب-والتقييم	نسبة الخطأ في التعرف الكلمي	سمات النظام المختصرة	ملاحظات
Menasri et al. ٢٠٠٧[٤٧]	abc–d	١٢,٦	• نظام هجين من: - نماذج ماركوف الخفية - والشبكات العصبية الاصطناعية • مبني على التقطيع الصريح	
Benouareth et al. ٢٠٠٨ [١٢]	abc–d	٩,٨٠	• نماذج ماركوف الخفية شبه-المتصلة - سلسلة بعدد ثابت من المراحل لكل محرف	
Schambach et al. ٢٠٠٨ [٤٤]	abcde–f	١٢,٧٨	• ثلاث نماذج ماركوف للتعرف على المحارف	النظام الفائز في ICDAR٢٠٠٧ [٧٦]
	abcde–s	٢٦,٠٦		

مرجع النظام	أجزاء القاعدة المستعملة للتدريب-والتقييم	نسبة الخطأ في التعرف الكلمي	سمات النظام المختصرة	ملاحظات
Al-Hajj et al. ٢٠٠٩[٢٩]	abc–d	٩,٠٤	• عدة نماذج ماركوف متصلة للتعرف على المحارف والفراغات البيضاء • تنفيذ فكرة النوافذ المنزلقة المائلة إضافة إلى العادية	أصحاب النظام الفائز آنفا في ICDAR٢٠٠٥ [٧٥]
Dreuw et al. ٢٠٠٨ and ٢٠٠٩[٢٥][٥٣]	abc-d	٥,٨٢	• نماذج ماركوف متصلة للتعرف على المحارف والفراغات البيضاء • تمت الاستعانة بتحويرات صورية لزيادة تنوع صور التدريب • إمكانية التأقلم على خط معين متاحة • الملامح مبنية على شرائح الصور • استخدام خوارزمية «تحليل المكونات الرئيسية» Principal component analysis (PCA) لتقليل عدد الملامح	
	abcd-e	١١,٢٢		

مرجع النظام	أجزاء القاعدة المستعملة للتدريب- والتقييم	نسبة الخطأ في التعرف الكلمي	سمات النظام المختصرة	ملاحظات
Kessentini et al. ٢٠١٠ [٣١]	abcd–e	٢٠,٤	• نماذج ماركوف متعددة الروافد • ملامح كنتورية ومن العناصر الصورية • كل ملمح يعبر في رافد مستقل	
	abcde–f	١٧,٩١		
	abcde–s	٢٥,٤٩		
Pechwitz et al. ٢٠١٢ [٤١]	abc–d	٨,٢	• نماذج ماركوف الخفية شبه-المتصلة للمحارف • ملامح من العناصر الصورية • التدريب بخوارزمية Viterbi	
Natarajan et al. ٢٠١٢ [٨]	abc–d	١٠,٦	• نماذج ماركوف الخفية المتصلة للمحارف • التأقلم التلقائي على خط معين	
Rothacker and Fink ٢٠١٢ [٨٦]	abc–d	٣,٨	• نماذج ماركوف مع حقيبة ملامح • استخدام «تحليل المكونات الرئيسية» (PCA) لتقليل عدد الملامح	
٢٠١٢ Graves [٦٠]	abcde–f	٦,٦٣	• شبكة عصبية متكررة (نواة تعلم عميق) • بنية ثنائية للذاكرة قصيرة المدى طويلة • ملامح من العناصر الصورية	صاحب النظام الفائز في مسابقة ICDAR ٢٠٠٩ [٨٧]
	abcde–s	١٨,٩٤		

مرجع النظام	أجزاء القاعدة المستعملة للتدريب-والتقييم	نسبة الخطأ في التعرف الكلمي	سمات النظام المختصرة	ملاحظات
Parvez and Mahmoud ٢٠١٣ [٨٨]	التدريب على حروف مقطعة ليست من IFN/ENIT والتقييم على abcd–e	٢٠,٤٢	• نظام هيكلي • مصنف الجار الأقرب • المحارف ممثلة عبر مقاربة المضلعات الضبابية	
Azeem and Ahmed ٢٠١٣ [٣٠]	abc–d	٢,٣	• عدة نماذج ماركوف متصلة للتعرف على المحارف والفراغات البيضاء بعد تطبيعه عرضها • ملامح التدرج والتقعر • إعادة تنفيذ فكرة النوافذ المنزلقة المائلة إضافة إلى العادية [٢٩]	
	abcd–e	٦,٥٦		
	abcde–f	٦,٩		
	abcde–s	١٥,٢		
Giménez et al. ٢٠١٤ [٩٠]	abc–d	٤,٧	• نماذج ماركوف الخفية البيرنولية • الملامح: العناصر الصورية الثنائية	أصحاب النظام الفائز [٨٩] في ICFHR٢٠١٠ [٧٨]
	abcd–e	٦,١		
	abcde–f	٧,٨٠		
	abcde–s	١٥,٣٨		
Abandah et al. ٢٠١٤ [٦١]	abc–d	١,٠٤	• شبكة عصبية متكررة (نواة تعلم عميق) • بنية ثنائية للذاكرة قصيرة المدى طويلة • تقطيع صريح • عدة ملامح متنوعة	
	abcd–e	٦,٥٤		
	abcde–f	٧,٥٤		
	abcde–s	١٥,٢٠		

مرجع النظام	أجزاء القاعدة المستعملة للتدريب-والتقييم	نسبة الخطأ في التعرف الكلمي	سمات النظام المختصرة	ملاحظات
Hamdani et al. ٢٠١٤ [٧٩] [٢٦]	abcde–f	٧,٨٠	• شبكة عصبية متكررة مع نماذج ماركوف الخفية المتصلة • الملامح تضمنت العناصر الصورية الرمادية • استخدام «تحليل المكونات الرئيسية» (PCA) لتقليل عدد الملامح • تم استخدام خوارزمية Viterbi جزئيا	عرضوا النظام الفائز في ICFHR ٢٠١١ [٧٩]
	abcde–s	١٥,٤٥		
Ahmad et al. ٢٠١٣. ٢٠١٤ [٤٩،٥٠]	abc–d	٢,٤٤	• نماذج ماركوف الخفية المتصلة متعددة الروافد • نماذج لأبعاض المحارف (تحت-المحرف أو -sub characters) وللفراغات البيضاء وللتطويل بين الحروف	
	abcd–e	٥,٥٥		
	abcde–f	٦,٤٠		
	abcde–s	١٢,١٤		
Ahmad and Fink [٥٢]	abc–d	١,٩٢	• نماذج ماركوف الخفية المتصلة متعددة الروافد • فصل الكتابة عن النقط والتشكيل	
	abcd–e	٥,٠٧		
	abcde–f	٧,٧٠		
	abcde–s	١٥,٤٥		

مرجع النظام	أجزاء القاعدة المستعملة للتدريب-والتقييم	نسبة الخطأ في التعرف الكلمي	سمات النظام المختصرة	ملاحظات
Stahlberg and Vogel ٢٠١٥ [٤٢]	abc–d	تبدأ من ٢,٤ وتزيد	• نماذج ماركوف الخفية مع التعلم العميق للتدريب • تقطيع ضمني باستخدام نماذج ماركوف الخفية تهيئة للتعلم العميق • الملامح تضمنت العناصر الصورية الرمادية • استخدام «تحليل المكونات الرئيسية» (PCA) لتقليل عدد الملامح • أقلمة التدريب لخط الكاتب المعين	اختلفت النتائج باختلاف تكوينات الملامح واستراتيجيات التدريب
	abcd–e	تبدأ من ٦,١ وتزيد		
	abcde–f	تبدأ من ٦,٨ وتزيد		
	abcde–s	تبدأ من ١١,٥ وتزيد		

٣,٢,٤ أهم بحوث التعرف الآلي على كلمات حرة

وأخيرا، نعرض في جدول ٤ نتائج أنظمة التعرف على نصوص الصور التي تحوي عدة كلمات حرة. فالفرق بين ما ههنا وما قبله أن هذه بمقدورها الاعتماد على نماذج لغوية عامة لتحسين النتائج. علاوة على ذلك، فبعض الأنظمة هنا تعالج صورا تتضمن عدة أسطر، مما يضفي بعدا آخر مهما للمسألة، وهو تقطيع الأسطر ضمنيا.

جدول ٤ بعض أهم بحوث التعرف على الصور التي تحوي عدة كلمات حرة.

مرجع النظام	نسبة الخطأ في التعرف الكلمي	قاعدة البيانات المستعملة	سمات النظام المختصرة
Saleem et al. ٢٠٠٩ [٩١]	٣٠,٠٪	قاعدة مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا للأحرف، وتشمل: • ٨,٢٥٠ وثيقة للتدريب • ٢١٨ وثيقة للتطوير • ٢٢٤ وثيقة للتقيين	• سلاسل ماركوف المتصلة • العديد من الملامح، وتم تقليص عددها آليا • الهدف: التعرف على المحارف ومن ثم الكلمات • استخدمت نماذج لغوية مداها ٣ أحرف قدرت من مدونة نصية قوامها ٩٠ مليون كلمة عربية (٩٢ ألف كلمة بحذف التكرار)
Natarajan et al. ٢٠١٢ [٨]	٢٥,٢٪	قاعدة «مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا» للأحرف، وتشمل: • ٣٧,٦٠٨ وثيقة للتدريب • ٨٦٨ وثيقة للتطوير • ٨٨٥ وثيقة للتقييم	• سلاسل ماركوف المتصلة • العديد من الملامح، وتم تقليص عددها آليا • الهدف: التعرف على المحارف ومن ثم الكلمات • إمكانية التأقلم على خط كاتب معين • استخدمت نماذج لغوية مداها ٣ أحرف قدرت من مدونة نصية قوامها ٢١٧ مليون كلمة عربية (١٢٠ ألف كلمة بحذف التكرار)

مرجع النظام	نسبة الخطأ في التعرف الكلمي	قاعدة البيانات المستعملة	سمات النظام المختصرة
Hamdani et al. ٢٠١٣ [٩٢]	٣٤,١٪ للتعرف المقيد على ٩٠ ألف كلمة بدون التكرار	مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا • ٤٢ ألف صفحة للتدريب • و٤٧٠ صفحة للتطوير	• سلاسل ماركوف المتصلة • الملامح تضمنت العناصر الصورية الرمادية • «تحليل المكونات الرئيسية» (PCA) لتقليل عدد الملامح • يستخدم التعرف المقيد كلمات التدريب كنموذج لغوي بينما يستخدم التعرف غير المقيد مدونة نصية من مليار كلمة تقريبا
	٢٥,٩٪ للتعرف بدون قيود على ٢٠٠ ألف كلمة بدون التكرار		
	٣٢,٥٪ للتعرف المقيد على ١٥ ألف كلمة بدون التكرار ٢٦,٨٪ للتعرف بدون قيود على ٢٠٠ ألف كلمة بدون التكرار	قاعدة بيانات «خط» • و٩,٤٧٥ سطرا للتدريب • و١,٩٠٢ سطرا للتطوير • و١,٩٩٧ سطرا للتقييم	

مرجع النظام	نسبة الخطأ في التعرف الكلمي	قاعدة البيانات المستعملة	سمات النظام المختصرة
Hamdani et al. ٢٠١٤ [٢٦]	٪٢٦,٨ للتعرف المقيد على ٩٤ ألف كلمة بدون التكرار ٪١٧,٠ للتعرف غير المقيد	قاعدة بيانات مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا • ٤٢ ألف صفحة للتدريب • و٤٧٠ صفحة للتطوير	• التعلم العميق (BLSTM مع RNNs) جنبا إلى جنب مع سلاسل ماركوف المتصلة • الملامح تضمنت العناصر الصورية الرمادية • «تحليل المكونات الرئيسية» (PCA) لتقليل عدد الملامح • يستخدم التعرف المقيد كلمات التدريب كنموذج لغوي بينما يستخدم التعرف غير المقيد مدونة نصية من مليار كلمة تقريبا • إمكانية التأقلم على خط معين
Hamdani et al. ٢٠١٤ [٦٢]	٪١٩,٩	قاعدة بيانات مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا • ٤٢ ألف صفحة للتدريب • و٤٧٠ صفحة للتطوير • و٦٣٣ صفحة للتقييم	• نظام هجين من التعلم العميق ونماذج ماركوف الخفية • الملامح تضمنت العناصر الصورية الرمادية • نموذج لغوي من ٤ أحرف محسوب من ٤٠٠ ألف كلمة (بدون التكرار محسوبة من مدونة نصية من مليار كلمة)

مرجع النظام	نسبة الخطأ في التعرف الكلمي	قاعدة البيانات المستعملة	سمات النظام المختصرة
Cao et al. ٢٠١٤ [٣٤]		قاعدة بيانات مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا	• سلاسل ماركوف المتصلة • عدة ملامح من ضمنها مرشحات «جابر» • إمكانية التأقلم على خط معين • تهجين عدة أنظمة لتحسين النتائج
	٤ , ٧٪	• مجموعةNIST OpenHaRT ٢٠١٣	
	١ , ٢٢٪	• المجموعة i	
Bluche et al. ٢٠١٤ [٦٤]	• ٢٠ , ١ للتعرف المقيد • ١٨ , ٤ للتعرف غير المقيد	قاعدة بيانات مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا	• نظام هجين من التعلم العميق ونماذج ماركوف الخفية للتعرف على المحارف • الملامح هي العناصر الصورية • يتم التعرف بعد ٤ مسوحات من الجهات الأربعة • التدريب على كلمات منعزلة ثم على أسطر • نموذج لغوي من ٣ أحرف محسوب من ٦٠ ألف كلمة • يستخدم التعرف المقيد كلمات التدريب كنموذج لغوي بينما يستخدم التعرف غير المقيد مدونة نصية من مليار كلمة تقريبا (GigaWord)

مرجع النظام	نسبة الخطأ في التعرف الكلمي	قاعدة البيانات المستعملة	سمات النظام المختصرة
Moysset et al. ٢٠١٤ [٦٣]	٢٩,٥٪	قاعدة بيانات مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا من ٩,٧٢٩ منطقة نصية • ١,٨٣٥ منطقة نصية للتدريب • ١,٥٨٢ منطقة نصية للتطوير	• نظام هجين من التعلم العميق ونماذج ماركوف الخفية للتعرف على المحارف، والكلمات، وأجزاء الكلمات العربية • الملامح هي العناصر الصورية • يتم التعرف بعد ٤ مسوحات من الجهات الأربعة • التدريب بدأ بالكلمات الأدق ثم الأقل دقة ثم بتحويرات صورية على الصور الأصلية • تضمن تقطيعا ضمنيا للأسطر

مرجع النظام	نسبة الخطأ في التعرف الكلمي	قاعدة البيانات المستعملة	سمات النظام المختصرة
BenZeghiba et al. ٢٠١٥ [٥٨]	٩, ٣٠٪ عند استعمال نماذج لغوية من ٤ أحرف لأجزاء الكلمات ٢, ٣٣٪ عند استعمال نماذج لغوية من ٣ أحرف للكلمات وأجزاء الكلمات	قاعدة بيانات من مشروع تحليل وترجمة وتصنيف المستندات متعددة اللغات آليا • ١٣, ٤٩٦ سطرا للتدريب • ١, ١٢٥ سطرا للتطوير • ٢, ٠٩٣ سطرا للتقييم	• نظام هجين من التعلم العميق ونماذج ماركوف الخفية • نماذج لغوية لأشهر الكلمات، وأجزاء الكلمات العربية
	٣, ٣١٪ عند استعمال نماذج لغوية من ٤ أحرف لأجزاء الكلمات ٢, ٣٣٪ عند استعمال نماذج لغوية من ٣ أحرف للكلمات وأجزاء الكلمات	قاعدة بيانات خط • ٤, ٤٢٨ سطرا للتدريب • ٨٧٦ سطرا للتطوير • ٩٥٩ سطرا للتقييم	

مرجع النظام	نسبة الخطأ في التعرف الكلمي	قاعدة البيانات المستعملة	سمات النظام المختصرة
Stahlberg and Vogel ٢٠١٥ [٤٢]	بين ٥,٣٠٪ و٣١,٦٪ حسب تعديلات النظام	قاعدة بيانات خط • ٩,٤٦٢ سطرا للتدريب • ١,٨٩٩ سطرا للتطوير • ١,٩٩٦ سطرا للتقييم	• التعلم العميق ونماذج ماركوف الخفية • الملامح تضمنت العناصر الصورية الرمادية • «تحليل المكونات الرئيسية» (PCA) لتقليل عدد الملامح • إمكانية التأقلم على خط كاتب معين • نموذج لغوي ثلاثي الأحرف مستنتج من بيانات التدريب في قاعدة بيانات «خط»

٥- أبرز أوعية النشر في مجال التعرف الآلي على النصوص المكتوبة

إن التعرف على النصوص المكتوبة -بما في ذلك التعرف على النص العربي- كما هو من فروع الذكاء الاصطناعي، فهو أحد تطبيقات مجال التعرف على الأنماط (Pattern Recognition)؛ لذا، فإن كثيرا من نشاطات المجال العلمية تقع ضمن اختصاصات الرابطة الدولية للتعرف على الأنماط (International Association for Pattern Recognition أو IAPR) وهي رابطة دولية تجمع المنظمات العلمية والمهنية غير الربحية ذات العلاقة، وهي تعتمد منظمةً واحدةً فقط من كل دولة يشارك عبرها الأفراد المهتمون بأنشطتها. وفيما يلي ثبت بأبرز المؤتمرات والمجلات المتعلقة بالرابطة المذكورة وبغيرها حيث يمكن نشر البحوث المتعلقة بالتعرف على النصوص العربية المكتوبة بخط اليد فيها، نقسمها إلى مؤتمرات ومجلات علمية.

١ , ٥ أهم مؤتمرات المجال الدولية

تنبع أهمية حضور المؤتمرات المتخصصة والنشر فيها من كونها بيئة مكثفة لتلاقح الأفكار وفرص النقاش والتعرف على أحدث النشاطات وأنشط الباحثين في المجال. كما أنها قد تشكل مسارت نشر سريعة للأفكار الجديدة، حيث لا تحتاج لنفس درجة التمحيص والإثباتات التي تشترطها المجلات. لذا، فقد ارتأينا إثراء الباب بنبذ عن بعض أهم المؤتمرات التي قد تهتم بمناقشة القراءة الآلية.

١ , ١ , ٥ المؤتمر الدولي لحدود التعرف على خط اليد

إن المؤتمر الدولي لحدود التعرف على خط اليد (International Conference on Frontiers in Handwriting Recognition أو ICFHR) مؤتمر رئيسي لبحوث وتطبيقات التعرف على خط اليد يجمع خبراء من الأوساط الأكاديمية والصناعية لتبادل الخبرات وتعزيز البحث المشترك وتطويره.

يوفر هذا المؤتمر ملتقى للباحثين في مجالات التعرف الفوري والمتراخي، وواجهات التعامل بالقلم، ومعالجة النماذج والاستبيانات آليا، ومكتبات الخط الرقمية، والوصول واستعادة مستندات الإنترنت. تتبنى الرابطة الدولية للتعرف على الأنماط هذا المؤتمر برعاية لجنتها الفنية (أنظمة القراءة)، إذ تقام فعاليات المؤتمر مرة كل عامين (للأعوام الزوجية)، وقد كان آخر انعقاد له عام ٢٠١٨ في منطقة شلالات نياغارا بالولايات المتحدة الأمريكية، وسيكون انعقاده القادم عام ٢٠٢٠ في دورتموند، ألمانيا -إن شاء الله-. يتم نشر البحوث المقبولة من قبل المؤتمر بواسطة معهد مهندسي الكهرباء والإلكترونيات (IEEE).

٢ , ١ , ٥ المؤتمر الدولي لتحليل الوثائق والتعرف عليها

ربما تعد سلسلة المؤتمرات الدولية لتحليل الوثائق والتعرف عليها (International Conference on Document Analysis and Recognition أو ICDAR) الأنجح في المجال، إذ هي أكبر تجمع دولي ورئيس للباحثين والعلماء والممارسين في مجتمع تحليل المستندات -بشكل عام-.

يوفر هذا المؤتمر منصة بارزة لمناقشة وتشجيع وتبادل الآراء حول أحدث التطورات في تحليل المستندات وفهمها واسترجاعها وتقييمها، حيث يشمل مصطلح «المستندات» عندهم أنواعا مختلفة من الوثائق: ابتداء من أوراق البردي التاريخية، ومرورا بالمستندات الورقية، إلى الصور الملتقطة بالكاميرا حتى المستندات الحديثة متعددة الوسائط.

تمت المصادقة على هذا المؤتمر من قبل اللجنة التقنية العاشرة للرابطة الدولية للتعرف على الأنماط IAPR (التعرف على الأشكال الرسومية) واللجنة التقنية الحادية عشرة (أنظمة القراءة)، وكان المؤتمر قد تأسس منذ ما يقرب من ثلاثة عقود، وهو يقام حاليا مرة كل عامين. عقد مؤتمرICDAR الأخير عام ٢٠١٧ في كيوتو باليابان. وسيعقد القادم عام ٢٠١٩ في سيدني بأستراليا -إن شاء الله تعالى-. يتم نشر بحوث المؤتمر وإصداراتهم بواسطة معهد مهندسي الكهرباء والإلكترونيات (IEEE).

٣,١,٥ ورشة العمل الدولية لأنظمة تحليل المستندات

ورشة العمل الدولية لأنظمة تحليل المستندات (International Workshop on Document Analysis Systems أو DAS) بوتقةٌ مهمة أيضا لبحوث التعرف على النصوص المكتوبة بخط اليد. تقام ورشات العمل هذه كل عامين، وقد عقدت آخرهن -حتى كتابة هذا الكتاب- عام ٢٠١٨ في فيينا في النمسا، وستعقد ورشة العمل التالية عام ٢٠٢٠ في ووهان في الصين -إن شاء الله-.

٤,١,٥ المؤتمر الدولي للتعرف على الأنماط

المؤتمر الدولي للتعرف على الأنماط (International Conference on Pattern Recognition أو ICPR) من أقدم المؤتمرات المرعية من قِبَل IAPR ومن أرسخها في مجال التعرف على الأنماط عموما. يرحب المؤتمر بالموضوعات المتعلقة بالتعرف على النصوص المكتوبة بخط اليد ضمن اهتماماته. ويعقد المؤتمر كل عامين. كان انعقاده الأخير (الرابع والعشرون) عام ٢٠١٨ في بكين بالصين، وسيعقد المؤتمر الدولي الخامس والعشرون-إن شاء الله- عام ٢٠٢٠ في ميلانو، إيطاليا.

٥,١,٥ الورشة الدولية لتحليل النصوص العربية ومشتقاتها والتعرّف الآلي عليها

الورشة الدولية لتحليل النصوص العربية ومشتقاتها والتعرّف الآلي عليها International Workshop on Arabic and Derived Script Analysis and) Recognition أو ASAR) بوتقة سنوية حديثة -نسبيا- متخصصة في تحليل النصوص العربية ونصوص اللغات المشتقة من العربية والتعرّف الآلي عليها.

تنعقد ورشة العمل الثالثة عام ٢٠١٩ بالتزامن مع انعقاد ICDAR للعام ٢٠١٩ في مدينة سيدني، أستراليا؛ وقد كانت ورشة العمل الثانية عام ٢٠١٨ في مدينة لندن، المملكة المتحدة؛ والأولى (عام ٢٠١٧) في نانسي، فرنسا.

وإضافة للمؤتمرات وورشات العمل المذكورة، تنعقد مؤتمرات أخرى ربما تكون ذات صلة ببعض مواضيع التعرف الآلي على الكتابة العربية، مثل المؤتمر الدولي للتعرف على الأنماط وذكاء الآلة (International Conference on Pattern Recognition and Machine Intelligence أو PReMI) والمؤتمر الدولي لتحليل ومعالجة الصور (International Conference on Image Analysis and Processing أو ICIAP)، والمؤتمر الدولي لتحليل الصور والأنماط الحاسوبية (International Conference on Computer Analysis of Images and Patterns أو CAIP) وحلقات العمل الدولية المشتركة مع IAPR حول التقنيات الإحصائية للتعرف على الأنماط (IAPR Joint International Workshops on Statistical Techniques in Pattern Recognition أو SPR) وكذلك التعرف على الأنماط الهيكلية والنحوية (Structural and Syntactic Pattern Recognition أو SSPR). ويمكن الاطلاع على قائمة المؤتمرات المعتمدة من IAPRفي صفحتهم على الشبكة العنكبوتية.

٥,٢ أهم المجلات العلمية المحكمة التي تصلح لنشر المقالات في المجال

نلقي فيما يلي بعض الضوء على بعض المجلات البارزة التي يتم فيها نشر البحوث المتعلقة بالتعرف على النصوص المكتوبة بخط اليد باللغة العربية:

١, ٢, ٥ المجلة الدولية لتحليل والتعرف على المستندات

تركز المجلة الدولية لتحليل والتعرف على المستندات (The International Journal on Document Analysis and Recognition أو IJDAR) على نشر المقالات العلمية المحكمة المتخصصة في تحليل الوثائق والتعرف عليها. يتضمن ذلك المساهمات التي تتناول التعرف على المحارف والأرقام والنصوص والخطوط والرسومات والصور والكتابة اليدوية والتوقيعات، بالإضافة إلى مجال تحليل هياكل الوثائق؛ كل ذلك بهدف فهم محتواها الدلالي آليا. تنشر البحوث المقبولة في هذه المجلة بواسطة الناشر Springer Verlag.

٢, ٢, ٥ تداولات معهد مهندسي الكهرباء والإلكترونيات لتحليل الأنماط والذكاء الآلي

لمعهد مهندسي الكهرباء والإلكترونيات الدولي (IEEE) عدة «تداولات» (Transactions) مهمة، منها رسائل تحليل الأنماط والذكاء الآلي (Transactions on Pattern Analysis and Machine Intelligence أو TPAMI). تعد هذه البوتقة من أشهر المجلات وأجودها في المجال، وهي تنشر في جميع المجالات التقليدية لرؤية الحاسب وفهم الصورة، وكذلك المجالات التقليدية لتحليل النماذج والتعرف عليها، ومجالات مختارة من ذكاء الآلة، مع التركيز على التعلم الآلي لتحليل الأنماط. كما يمكن أحيانا تغطية تقنيات البحث المرئي، وتحليل المستندات والخط اليدوي، وتحليل الصور الطبية، وتحليل الفيديو وغيرها. تصدر المجلة ١٢ عددًا في السنة.

٣, ٢, ٥ التعرف على الأنماط

التعرف على الأنماط (Pattern recognition أو PR) بوتقة مهمة أخرى في المجال. أنشئت المجلة منذ ما يقارب ٥٠ عامًا، -أي في السنوات الأولى لتطور علوم الحاسب الآلي ثم توسعت بشكل أكبر.

تقبل المجلة الأوراق التي تقدم مساهمات أصيلة في نظريات ومنهجيات وتطبيقات التعرف على الأنماط في أي مجال، بشرط أن يتم شرح سياق العمل بشكل واضح وترسيخه في أدبيات التعرف على الأنماط. تنشر المجلة ١٢ عددا في العام 12 عددًا في السنة بواسطة Elsevier Science B.V.

٤,٢,٥ رسائل التعرف على الأنماط

مجلة «رسائل التعرف على الأنماط» (Pattern Recognition Letters أو PRL) المحكّمة تنشر مقالات موجزة بوقت سريع (نسبيا) بتغطية واسعة لأدبيات التعرف على الأنماط (وخصوصا المواضيع التي تهتم بها كل من اللجان الفنية لمعهد الرابطة الدولية للتعرف على الأنماط)، تقبل المجلة الأوراق البحثية النظرية والمنهجية والتجريبية والتطبيقية. معايير قبول المقالات تتركز في أصالة البحث وجودته ووضوح طرحه. يتم نشر المجلة شهريًا بواسطة Elsevier Science B.V.

٦- الخاتمة

قطعت القراءة الآلية أشواطا منذ ظهرت، وما زالت معالجة الكتابة العربية تتطور في هذا المضمار مع أساليب تعلم الآلة الحديثة، خاصة ما لا يتطلب منها تقطيع الكلمات إلى حروف، كالتعرف الكلي والضمني في نماذج ماركوف الخفية والتعلم العميق. فصّل هذا الباب في شرح ومقارنة أحدث بحوث المجال، ثم خُتم بثبت لأهم مظان المراجع وأوعية النشر من مجلات ومؤتمرات، نسأل الله تعالى أن ينفع به قارئه وكاتبيه وناشره.

المراجع

[1] T. Gustav. Reading machine. US Patent 2.115.563. 1938.

[2] Timeline of optical character recognition. (n.d.). https://en.wikipedia.org/wiki/Timeline_of_optical_character_recognition.

[٣] سامح عويضة، قواعد البيانات الإلكترونية للمخطوطات التراثية العربية والإسلامية: الحاضر والمستقبل، الحرف العربي والتقنية، تحرير: يوسف العريان، مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ٢٠١٥.

[4] M.Z. Khedher. G.A. Abandah. Arabic character recognition using approximate stroke sequence. in: Third Int'l Conf. Lang. Resour. Eval. (LREC 2002). Canary Islands. Spain. 2002: pp. 28–34.

[5] Y. Elarian. S. Awaida. S.A. Mahmoud. Design of Datasets for Handwritten Arabic Texts Research. in: 1st Saudi High. Educ. Students Conf. Riyadh. 2010.

[6] S.A. Mahmoud. I. Ahmad. M. Alshayeb. W.G. Al-Khatib. M.T. Parvez. G.A. Fink. V. Märgner. H. EL Abed. KHATT: Arabic Offline Handwritten Text Database. in: Proc. 13th Int. Conf. Front. Handwrit. Recognit. (ICFHR 2012). IEEE. 2012: pp. 447-452.

[7] Y. Al-Ohali. M. Cheriet. C.Y. Suen. Databases for recognition of handwritten Arabic cheques. Pattern Recognit. 36 (2003) 111–121. doi:10.1016/S0031-3203(02)00064-X.

[8] P. Natarajan. R. Prasad. H. Cao. K. Subramanian. S. Saleem. D. Belanger. S. Vitaladevuni. M. Kamali. E. MacRostie. Arabic Text Recognition Using a Script-Independent Methodology: A Unified HMM-Based Approach for Machine-Printed and

Handwritten Text. in: V. Märgner. H. El Abed (Eds.). Guid. to OCR Arab. Scripts. Springer London. London. 2012: pp. 485-505. doi:10.1007/978-1-4471-4072-6_20.

[9] Y. Elarian. I. Ahmad. S. Awaida. W. Al-Khatib. A. Zidouri. Arabic ligatures: Analysis and application in text recognition. in: Proc. Int. Conf. Doc. Anal. Recognition. ICDAR. 2015. doi:10.1109/ICDAR.2015.7333891.

[10] Y. Elarian. A Lexicon of Connected Components for Arabic Optical Text Recognition. Jordan University of Science and Technology. Irbid. Jordan. 2006.

[11] U. V. Marti. H. Bunke. The IAM-database: An English sentence database for offline handwriting recognition. Int. J. Doc. Anal. Recognit. 5 (2003) 39–46. doi:10.1007/s100320200071.

[12] A. Benouareth. A. Ennaji. M. Sellami. Semi-continuous HMMs with explicit state duration for unconstrained Arabic word modeling and recognition. Pattern Recognit. Lett. 29 (2008) 1742–1752.

[13] M. Pechwitz. V. Märgner. H. El Abed. Comparison of Two Different Feature Sets for Offline Recognition of Handwritten Arabic Words. Proc. Tenth Int. Work. Front. Handwrit. Recognit. (IWFHR 2006). (2006). https://hal.archives-ouvertes.fr/inria-00112643/ (accessed February 9. 2016).

[14] G.A. Abandah. F.T. Jamour. Recognizing handwritten Arabic script through efficient skeleton-based grapheme segmentation algorithm. in: 2010 10th Int. Conf. Intell. Syst. Des. Appl.. 2010: pp. 977–982.

[15] A.M. Al-Shatnawi. K. Omar. A comparative study between methods of Arabic baseline detection. in: Proc. Int. Conf. Electr. Eng. Informatics. 2009: pp. 73–77. doi:10.1109/ICEEI.2009.5254814.

[16] H. El Abed. V. Märgner. Comparison of Different Preprocessing and Feature Extraction Methods for Offline Recognition of Handwritten ArabicWords. in: Proc. Ninth Int. Conf. Doc. Anal. Recognit. (ICDAR 2007). 2007: pp. 974-978. doi:10.1109/ICDAR.2007.4377060.

[17] Text extraction from skew images opencv. (n.d.). https://stackoverflow.com/questions/34022113/text-extraction-from-skew-images-opencv.

[18] H. Akram. S. Khalid. others. Using features of local densities. statistics and HMM toolkit (HTK) for offline Arabic handwriting text recognition. J. Electr. Syst. Inf. Technol. 4 (2017) 387–396.

[19] A.M. Al-Shatnawi. A Preprocessing Model For Handwritten Arabic Texts Based on Voronoi Diagrams. Int. J. Comput. Sci. Inf. Technol. 7 (2015). doi:10.5121/ijcsit.2015.7601.

[20] M. Wienecke. G.A. Fink. G. Sagerer. Toward automatic video-based whiteboard reading. Int. J. Doc. Anal. Recognit. 7 (2005) 188–200.

[21] Y. Elarian. Analysis of Some Arabic Scripting Units in Computational-Linguistic Resources. in: 1st Saudi High. Educ. Students Conf. Riyadh. 2010.

[22] Y.S. Elarian. S.A. Mahmoud. An Adaptive Line Segmentation Algorithm (ALSA) for Arabic. in: Proc. Int. Conf. Comput. Vis. Pattern Recognit.. 2008: pp. 735–739.

[23] Y. Elarian. A. Zidouri. W. Al-Khatib. Ground-Truth and Metric for the Evaluation of Arabic Handwritten Character Segmentation. in: 2014 14th Int. Conf. Front. Handwrit. Recognit.. 2014: pp. 766–770.

[24] I.S. Abuhaiba. A discrete Arabic script for better automatic document understanding. Arab. J. Sci. Eng. 28 (2003) 77–94.

[25] P. Dreuw. S. Jonas. H. Ney. White-space models for offline Arabic handwriting recognition. in: Proc. 19th Int. Conf. Pattern Recognit. (ICPR 2008). 2008: pp. 1–4.

[26] M. Hamdani. P. Doetsch. M. Kozielski. A.E.-D. Mousa. H. Ney. The RWTH Large Vocabulary Arabic Handwriting Recognition System. in: Proc. 11th IAPR Int. Work. Doc. Anal. Syst. (DAS 2014). IEEE. 2014: pp. 111–115. doi:10.1109/DAS.2014.61.

[27] H. El Abed. V. Märgner. How to Improve a Handwriting Recognition System. in: Proc. 10th Int. Conf. Doc. Anal. Recognit. (ICDAR 2009). IEEE. 2009: pp. 1181-1185. doi:10.1109/ICDAR.2009.11.

[28] R. El-Hajj. L. Likforman-Sulem. C. Mokbel. Arabic handwriting recognition using baseline dependant features and hidden markov modeling. in: Proc. Eighth Int. Conf. Doc. Anal. Recognit. (ICDAR 2005). 2005: pp. 893–897.

[29] R. Al-Hajj Mohamad. L. Likforman-Sulem. C. Mokbel. Combining slanted-frame classifiers for improved HMM-based Arabic handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 1165–1177.

[30] S. Azeem. H. Ahmed. Effective technique for the recognition of offline Arabic handwritten words using hidden Markov models. Int. J. Doc. Anal. Recognit. 16 (2013) 399–412. doi:10.1007/s10032-013-0201-8.

[31] Y. Kessentini. T. Paquet. A.M. Ben Hamadou. Off-line handwritten word recognition using multi-stream hidden Markov models. Pattern Recognit. Lett. 31 (2010) 60–70.

[32] M. Dehghan. K. Faez. M. Ahmadi. M. Shridhar. Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM. Pattern Recognit. 34 (2001) 1057–1065. doi:10.1016/S0031-3203(00)00051-0.

[33] R. Safabakhsh. P. Adibi. Nastaaligh handwritten word recognition using a continuous-density variable-duration HMM. Arab. J. Sci. Eng. 30 (2005) 95–118.

[34] H. Cao. P. Natarajan. X. Peng. K. Subramanian. D. Belanger. N. Li. Progress in the Raytheon BBN Arabic Offline Handwriting Recognition System. in: Proc. Int. Conf. Front. Handwrit. Recognit. (ICFHR 2014). IEEE. 2014: pp. 555–560. doi:10.1109/ICFHR.2014.99.

[35] N. Azizi. N. Farah. M. Sellami. A. Ennaji. Using Diversity in Classifier Set Selection for Arabic Handwritten Recognition. in: N. Gayar. J. Kittler. F. Roli (Eds.). Proc. 9th Int. Work. Mult. Classif. Syst.. Springer Berlin Heidelberg. Berlin. Heidelberg. 2010: pp. 235–244. doi:10.1007/978-3-642-12127-2_24.

[36] B. Yanikoglu. P.A. Sandon. Segmentation of off-line cursive handwriting using linear programming. Pattern Recognit. 31 (1998) 1825–1833.

[37] Y. Elarian. F. Idris. A Lexicon of Connected Components for Arabic Optical Character Recognition. in: Int. Work. Front. Arab. Handwrit. Recognition. Istanbul. 2011.

[38] S. Alansary. M. Nagi. N. Adly. Processing Arabic Text Content: The Encoding Component in an Interlingual System for Man-Machine Communication in Natural Language". in: Proc. 6th Int. Conf. Lang. Eng.. 2006.

[39] S.A. Mahmoud. Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models. Signal Processing. 88 (2008) 844–857.

[40] S.M. Awaida. S.A. Mahmoud. A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models. Signal Processing. 89 (2009) 1176–1184.

[41] M. Pechwitz. H. El Abed. V. Märgner. Handwritten Arabic Word Recognition Using the IFN/ENIT-database. in: V. Märgner. H. El Abed (Eds.). Guid. to OCR Arab. Scripts. Springer London. 2012: pp. 297-313. doi:10.1007/978-1-4471-4072-6{_}8.

[42] F. Stahlberg. S. Vogel. The QCRI Recognition System for Handwritten Arabic. in: V. Murino. E. Puppo (Eds.). Proc. 18th Int. Conf. Image Anal. Process. (ICIAP 2015). Springer International Publishing. Genoa. Italy. 2015: pp. 276–286. doi:10.1007/978-3-319-23234-8_26.

[43] E. Chammas. C. Mokbel. L. Likforman-Sulem. Arabic handwritten document preprocessing and recognition. in: Proc. 13th Int. Conf. Doc. Anal. Recognit. (ICDAR 2015). 2015: pp. 451–455. doi:10.1109/ICDAR.2015.7333802.

[44] M.P. Schambach. J. Rottland. T. Alary. How to convert a Latin handwriting recognition system to Arabic. in: Proc. 11th Int. Conf. Front. Handwrit. Recognit. (ICFHR 2008). 2008: pp. 265–270.

[45] R. Al-Hajj Mohamad. C. Mokbel. L. Likforman-Sulem. Combination of hmm-based classifiers for the recognition of arabic handwritten words. in: Proc. Ninth Int. Conf. Doc. Anal. Recognit. (ICDAR 2007). 2007: pp. 959–963.

[46] M.S. Khorsheed. Recognising handwritten Arabic manuscripts using a single hidden Markov model. Pattern Recognit. Lett. 24 (2003) 2235–2242.

[47] F. Menasri. N. Vincent. E. Augustin. M. Cheriet. Shape-based alphabet for off-line Arabic handwriting recognition. in: Proc. Ninth Int. Conf. Doc. Anal. Recognit. (ICDAR 2007). 2007: pp. 969–973.

[48] M. Hamdani. H. El Abed. M. Kherallah. A.M. Alimi. Combining multiple HMMs using on-line and off-line features for off-line arabic handwriting recognition. in: Proc. 10th Int. Conf. Doc. Anal. Recognit. (ICDAR 2009). Ieee. 2009: pp. 201–205. doi:10.1109/ICDAR.2009.40.

[49] I. Ahmad. L. Rothacker. G.A. Fink. S.A. Mahmoud. Novel sub-character HMM models for arabic text recognition. in: Proc. Int. Conf. Doc. Anal. Recognition. ICDAR. 2013. doi:10.1109/ICDAR.2013.135.

[50] I. Ahmad. G.A. Fink. S.A. Mahmoud. Improvements in Sub-character HMM Model Based Arabic Text Recognition. in: Proc. 14th Int. Conf. Front. Handwrit. Recognit. (ICFHR 2014). IEEE. Crete. 2014: pp. 537–542. doi:10.1109/ICFHR.2014.96.

[51] Y.S. Elarian. I. Ahmad. S.M. Awaida. W.G. Al-Khatib. A. Zidouri. Arabic Ligatures: Analysis and Application in Text Recognition. in: Proc. 13th Int. Conf. Doc. Anal. Recognit. (ICDAR 2015). IEEE. 2015: pp. 896–900.

[52] I. Ahmad. G.A. Fink. Multi-stage HMM based Arabic text recognition with rescoring. in: Proc. 13th Int. Conf. Doc. Anal. Recognit. (ICDAR 2015). IEEE. 2015: pp. 751–755. doi:10.1109/ICDAR.2015.7333862.

[53] P. Dreuw. D. Rybach. C. Gollan. H. Ney. Writer Adaptive Training and Writing Variant Model Refinement for Offline Arabic Handwriting Recognition. in: Proc. 10th Int. Conf. Doc. Anal. Recognit. (ICDAR 2009). IEEE. 2009: pp. 21–25. doi:10.1109/ICDAR.2009.9.

[54] A. Benouareth. A. Ennaji. M. Sellami. HMMs with Explicit State Duration Applied to Handwritten Arabic Word Recognition. in: Proc. 18th Int. Conf. Pattern Recognit. (ICPR 2006). IEEE. 2006: pp. 897–900. doi:10.1109/ICPR.2006.631.

[55] S. Alma'adeed. C. Higgins. D. Elliman. Recognition of off-line handwritten Arabic words using hidden Markov model approach. in: Proc. Object Recognit. Support. by User Interact. Serv. Robot.. IEEE Comput. Soc. 2002: pp. 481–484. doi:10.1109/ICPR.2002.1047981.

[٥٦] يوسف العريان (محررا)، الحرف العربي والتقنية، مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ٢٠١٥.

[57] P. Natarajan. D. Belanger. R. Prasad. M. Kamali. K. Subramanian. P. Natarajan. Baseline Dependent Percentile Features for Offline Arabic Handwriting Recognition. in: Proc. 11th Int. Conf. Doc. Anal. Recognit. (ICDAR 2011). IEEE. 2011: pp. 329–333. doi:10.1109/ICDAR.2011.74.

[58] M.F. BenZeghiba. J. Louradour. C. Kermorvant. Hybrid word/Part-of-Arabic-Word Language Models for arabic text document recognition. in: Proc. 13th Int. Conf. Doc. Anal. Recognit. (ICDAR 2015). IEEE. 2015: pp. 671–675. doi:10.1109/ICDAR.2015.7333846.

[59] A. Graves. J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. in: Adv. Neural Inf. Process. Syst.. 2009: pp. 545–552.

[60] A. Graves. Offline Arabic Handwriting Recognition with Multidimensional Recurrent Neural Networks. in: V. Märgner. H. El Abed (Eds.). Guid. to OCR Arab. Scripts. Springer London. London. 2012: pp. 297-313. doi:10.1007/978-1-4471-4072-6_12.

[61] G.A. Abandah. F.T. Jamour. E.A. Qaralleh. Recognizing handwritten Arabic words using grapheme segmentation and recurrent neural networks. Int. J. Doc. Anal. Recognit. 17 (2014) 275–291. doi:10.1007/s10032-014-0218-7.

[62] M. Hamdani. P. Doetsch. H. Ney. Improvement of Context Dependent Modeling for Arabic Handwriting Recognition. in: Proc. 14th Int. Conf. Front. Handwrit. Recognit. (ICFHR 2014). IEEE. 2014: pp. 494–499. doi:10.1109/ICFHR.2014.89.

[63] B. Moysset. T. Bluche. M. Knibbe. M.F. Benzeghiba. R. Messina. J. Louradour. C. Kermorvant. The A2iA Multi-lingual Text Recognition System at the Second Maurdor Evaluation. in: Proc. 14th Int. Conf. Front. Handwrit. Recognit. (ICFHR 2014). IEEE. 2014: pp. 297–302. doi:10.1109/ICFHR.2014.57.

[64] T. Bluche. J. Louradour. M. Knibbe. B. Moysset. M.F. Benzeghiba. C. Kermorvant. The A2iA Arabic Handwritten Text Recognition System at the Open HaRT2013 Evaluation. in: Proc. 11th IAPR Int. Work. Doc. Anal. Syst. (DAS 2014). IEEE. 2014: pp. 161–165. doi:10.1109/DAS.2014.40.

[65] O. Morillot. C. Oprean. L. Likforman-Sulem. C. Mokbel. E. Chammas. E. Grosicki. The UOB-Telecom ParisTech Arabic Handwriting Recognition and Translation Systems for the OpenHart 2013 Competition. in: Proc. 12th Int. Conf. Doc. Anal. Recognit. (ICDAR 2013). Washington DC. United States. 2013: p. NIST. https://hal.archives-ouvertes.fr/hal-00948985.

[66] T. Bluche. H. Ney. C. Kermorvant. A Comparison of Sequence-Trained Deep Neural Networks and Recurrent Neural Networks Optical Modeling for Handwriting Recognition. in: L. Besacier. A.-H. Dediu. C. Mart\'in-Vide (Eds.). Proc. Second Int. Conf. Stat. Lang. Speech Process. SLSP2014. Springer International Publishing. Grenoble. 2014: pp. 199–210. doi:10.1007/978-3-319-11397-5_15.

[67] N. Kharma. M. Ahmed. R. Ward. A New Comprehensive Database of Hadritten Arabic Words . Numbers . and Signatures used for OCR Testing. Can. Conf. Electr. Comput. Eng. (1999) 766–768.

[68] S. Al-Maadeed. D. Elliman. C. Higgins. A data base for Arabic handwritten text recognition research. in: Proc. Eighth Int. Work. Front. Handwrit. Recognit. (IWFHR 2002). IEEE Comput. Soc. 2002: pp. 485–489. doi:10.1109/IWFHR.2002.1030957.

[69] H. Alamri. J. Sadri. C.Y. Suen. N. Nobile. A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition Huda Alamri. in: Elev. Int. Conf. Front. Handwrit. Recognit.. Montreal. Canada. 2008.

[70] E. El-Sherif. S. Abdleazeem. A two-stage system for Arabic handwritten digit recognition tested on a new large database. in: Int. Conf. Artificial Intell. Pattern Recognit.. 2007: pp. 237–242.

[71] S.M. Strassel. Linguistic Resources for Arabic Handwriting Recognition. in: MEDAR Second Int. Conf. Arab. Lang. Resour. Tools. Cairo. Egypt. April 22-23. 2009: pp. 37–41.

[72] A. Tong. M. Przybocki. V. Märgner. H. El Abed. NIST 2013 Open Handwriting Recognition and Translation (Open HaRT-13) Evaluation. in: Proc. 11th IAPR Int. Work. Doc. Anal. Syst. (DAS 2014). IEEE. 2014: pp. 81-85. doi:10.1109/DAS.2014.43.

[73] NIST. OpenHaRT 2013 Information Page. (n.d.). http://www.nist.gov/itl/iad/mig/hart2013.cfm (accessed February 25. 2016).

[74] M. Pechwitz. S.S. Maddouri. V. Märgner. N. Ellouze. H. Amiri. IFN/ENIT - Database of Handwritten Arabic Words. in: 7th Colloq. Int. Francoph. Sur l-Ecrit Le Doc. . CIFED 2002. Hammamet. Tunis. 2002: pp. 129--136.

[75] V. Märgner. M. Pechwitz. H. El Abed. ICDAR 2005 Arabic handwriting recognition competition. in: Proc. Eighth Int. Conf. Doc. Anal. Recognit. (ICDAR 2005). IEEE. 2005: pp. 70-74 Vol. 1. doi:10.1109/ICDAR.2005.52.

[76] V. Märgner. H. El Abed. Arabic Handwriting Recognition Competition. in: Proc. Ninth Int. Conf. Doc. Anal. Recognit. (ICDAR 2007) Vol 2. IEEE. 2007: pp. 1274-1278. doi:10.1109/ICDAR.2007.4377120.

[77] H. El Abed. V. Märgner. ICDAR 2009-Arabic handwriting recognition competition. Int. J. Doc. Anal. Recognit. 14 (2010) 3-13. doi:10.1007/s10032-010-0117-5.

[78] V. Märgner. H. El Abed. ICFHR 2010 - Arabic Handwriting Recognition Competition. in: Proc. 12th Int. Conf. Front. Handwrit. Recognit. (ICFHR 2010). IEEE. 2010: pp. 709-714. doi:10.1109/ICFHR.2010.115.

[79] V. Märgner. H. El Abed. ICDAR 2011 - Arabic Handwriting Recognition Competition. in: Proc. 11th Int. Conf. Doc. Anal. Recognit. (ICDAR 2011). IEEE. 2011: pp. 1444-1448. doi:10.1109/ICDAR.2011.287.

[80] S.A. Mahmoud. I. Ahmad. W.G. Al-Khatib. M. Alshayeb. M. Tanvir Parvez. V. Märgner. G.A. Fink. KHATT: An open Arabic offline handwritten text database. Pattern Recognit. 47 (2014) 1096-1112. doi:10.1016/j.patcog.2013.08.009.

[81] H. Alamri. C. He. C.Y. Suen. A New Approach for Segmentation and Recognition of Arabic Handwritten Touching Numeral Pairs. Comput. Anal. Images Patterns. 5702 (2009) 165–172. doi:10.1007/978-3-642-03767-2.

[82] S.A. Mahmoud. W.G. Al-Khatib. Recognition of Arabic (Indian) bank check digits using log-gabor filters. Appl. Intell. 35 (2010) 445–456. doi:10.1007/s10489-010-0235-2.

[83] M. Cheriet. Y. Al-Ohali. N. Ayat. C.Y. Suen. Arabic Cheque Processing System: Issues and Future Trends. in: B.B. Chaudhuri (Ed.). Digit. Doc. Process.. Springer London. London. 2007: pp. 213–234. doi:10.1007/978-1-84628-726-8.

[84] S. Alma'adeed. C. Higgins. D. Elliman. Off-line recognition of handwritten Arabic words using multiple hidden Markov models. Knowledge-Based Syst. 17 (2004) 75–79. doi:http://dx.doi.org/10.1016/j.knosys.2004.03.002.

[85] N. Farah. L. Souici-Meslati. M. Sellami. Classifiers combination and syntax analysis for Arabic literal amount recognition. Eng. Appl. Artif. Intell. 19 (2006) 29–39. doi:10.1016/j.engappai.2005.05.005.

[86] L. Rothacker. S. Vajda. G.A. Fink. Bag-of-Features Representations for Offline Handwriting Recognition Applied to Arabic Script. in: Proc. 13th Int. Conf. Front. Handwrit. Recognit. (ICFHR 2012). 2012: pp. 149–154. doi:10.1109/ICFHR.2012.185.

[87] S. Mozaffari. H. Soltanizadeh. ICDAR 2009 Handwritten Farsi/Arabic Character Recognition Competition. in: Proc. 10th Int. Conf. Doc. Anal. Recognit. (ICDAR 2009). 2009: pp. 1413–1417. doi:10.1109/ICDAR.2009.283.

[88] M.T. Parvez. S.A. Mahmoud. Arabic handwriting recognition using structural and syntactic pattern attributes. Pattern Recognit. 46 (2013) 141–154. doi:10.1016/j.patcog.2012.07.012.

[89] A. Giménez. I. Khoury. A. Juan. Windowed Bernoulli Mixture HMMs for Arabic Handwritten Word Recognition. in: Proc. 12th Int. Conf. Front. Handwrit. Recognit. (ICFHR 2010). IEEE. 2010: pp. 533-538. doi:10.1109/ICFHR.2010.88.

[90] A. Giménez. I. Khoury. J. Andrés-Ferrer. A. Juan. Handwriting word recognition using windowed Bernoulli HMMs. Pattern Recognit. Lett. 35 (2014) 149-156. doi:10.1016/j.patrec.2012.09.002.

[91] S. Saleem. H. Cao. K. Subramanian. M. Kamali. R. Prasad. P. Natarajan. Improvements in BBN's HMM-Based Offline Arabic Handwriting Recognition System. in: Proc. 10th Int. Conf. Doc. Anal. Recognit. (ICDAR 2009). IEEE. 2009: pp. 773–777. doi:10.1109/ICDAR.2009.282.

[92] M. Hamdani. A.E.-D. Mousa. H. Ney. Open Vocabulary Arabic Handwriting Recognition Using Morphological Decomposition. in: Proc. 12th Int. Conf. Doc. Anal. Recognit. (ICDAR 2013). IEEE. 2013: pp. 280–284. doi:10.1109/ICDAR.2013.63.

الباب الثاني

التعرف الآلي على الكلام العربي المنطوق وتطبيقاته في القرآن الكريم

د. أحمد حمدي أبو عبسة

التعرف الآلي على الكلام العربي المنطوق وتطبيقاته في القرآن الكريم

د. أحمد حمدي أبو عبسة[١]

ملخص

يعتبر التعرف على الكلام العربي المنطوق من الأبحاث الهامة التي لها دور كبير في كثير من مجالات الحياة مثل التعليم والصحة والصناعة وغيرها من التطبيقات. في هذا البحث تم تطوير التعرف الآلي على الكلام العربي المنطوق باستخدام أنظمة الذكاء الاصطناعي وتطبيقه لخدمة القرآن الكريم. في الطريقة التقليدية لمعالجة الكلام، يتم الاعتماد على تقسيم الجملة الصوتية إلى مجموعة ثابتة من الأُطر، بينما في هذا البحث تم الاعتماد على المقاطع الصوتية والتي تعرف على أنها الجزء الأساسي الأصغر في اللغة والمكونة من مقاطع ساكنة ومقاطع متحركة. استعرضنا في هذا البحث أهم خوارزميات استخراج خصائص المقاطع الصوتية، والتي تعتبر الخطوة الأولى في تصنيف المقاطع الصوتية. ثم، لتحسين نتيجة التصنيف، قلصنا حجم مصفوفة خصائص المقاطع الصوتية باستخدام تقنية تحليل المكونات الأساسية. كما تم استخدام نظام التشجير التصنيفي المبني على قواعد التجويد، حيث يتم تصنيف المقاطع الصوتية إلى ثلاث مراحل: تصنيف نهاية المقطع الصوتي ساكن أم متحرك، وتصنيف الحرف الساكن من حيث التفخيم والغنة، وتحديد زمن الحرف المتحرك في المقطع الصوتي. من خلال تطبيق هذا البحث، وجدنا أن إدخال قواعد التجويد القرآنية مع أنظمة التعرف الآلي على الكلام لها دور مهم في تحسين دقة تصنيف البيانات القرآنية.

١- د.أحمد حمدي أبو عبسة رئيس قسم هندسة البرمجيات في جامعة فلسطين. حصل د. أبو عبسة على درجة البكالوريوس في هندسة الاتصالات والتحكم من الجامعة الإسلامية بغزة وعلى درجة الماجستير في علوم الحاسب الآلي من جامعة شمال فرجينيا ثم على ماجستير آخر في أنظمة الاتصالات من الجامعة الإسلامية بغزة. حصل على درجة الدكتوراة في معالجة الإشارة الرقمية من قسم الهندسة الكهربائية في جامعة الملك فهد للبترول والمعادن، وله العديد من الأبحاث والمشاريع في مجال معالجة الصوت والصورة بتقنيات الذكاء الاصطناعي.

١- مقدمة

بدأ اهتمام خبراء الحاسب والباحثين في مجال التعرف الآلي على الكلام منذ أكثر من أربعة عقود، وذلك لكي يصل الإنسان إلى مرحلة تجعله قادرا على التخاطب مع الحاسب الآلي وإعطائه الأوامر بدون الحاجة إلى الكتابة مما من شأنه توفير الجهد والوقت وإمكانية التفاعل مع الآلة بشكل طبيعي أكثر والتي تؤدي إلى استخدامها في مجالات تطبيقية متعددة.

ومع تطور التقنيات التكنولوجية في العصر الحديث، اتجه العالم إلى استخدام مفهوم الذكاء الاصطناعي (Artificial Intelligence أو AI) وتعلم الآلة (Machine Learning) في مجالات متعددة والتي من ضمنها التعرف الآلي على الكلام المنطوق باللغة العربية، وكذلك مجالات التعرف على أحكام التجويد في تلاوة القرآن الكريم.

يعرف مصطلح الذكاء الاصطناعي على أنه قدرة الآلة على محاكاة العقل البشري والتعلم من التجارب السابقة. ومنذ التطوّر الذي شهده الحاسب الآلي في منتصف القرن العشرين، تمكن العلماء من برمجة الحاسب الآلي وتطويره للقيام بمهام كثيرة ومعقدة تضاهي مستوى أداء الخبراء والمحترفين في مجالات كالتشخيص الطبي، أو في محركات البحث أو في تطبيقات التعرف على الصوت والكتابة اليدوية وغير ذلك [١].

ويمكن تقسيم أهداف الذكاء الاصطناعي وتعلم الآلة إلى ثلاثة أقسام رئيسية على النحو التالي:

١. التصنيف (Classification): حيث تقوم الخوارزمية بالتعلم وذلك من خلال وجود مجموعة من الأصناف Classes وكل صنف له خصائص features مشتركة، حيث يقوم المصنف بربط الخصائص بصنف معين.

٢. الارتباط (Regression): وهو أسلوب إحصائي يستخدم في قياس مدى العلاقة الدلالية بين متغيرين ، بحيث يكون أحد المتغيرات (متغير تابع) والآخر (متغير مستقل أو مُفسِر) وهو المتسبب في تغير المتغير التابع، وقد يستعمل للتنبؤ بقيم المتغير التابع بناء على المستقل.

٣. التجميع (Clustering): حيث تقوم الخوارزمية بتقسيم البيانات إلى مجموعات غير معروفة مسبقًا وكل مجموعة يتم التعامل معها على أنها صنف.

يُعرّف التعرف التلقائي على الكلام (ASR)Automatic Speech Recognition بأنه عملية تحويل الموجات الصوتية (الإشارات الصوتية للكلام) إلى كلمات أو وحدات لغوية Phonemes [٢]. يظهر التعرف التلقائي على الكلام في العديد من المجالات الصناعية والمدنية، بما في ذلك: التطبيقات التي تشجع الاستغناء عن احتياج الأيدي في التعامل معها، والتفاعل مع الأجهزة الذكية، والترجمة الشفوية التلقائية، وأدوات دعم المعاقين سمعيا، والإملاء التلقائي وغيرها من التطبيقات.

وعند تطبيق نظام التعرف الآلي على الكلام الصوتي في الحاسب الآلي، وجد أنه من السهل التعرف على الكلمات المنفردة، ولكن الأصعب هو التعرف على الكلام المستمر. وهذا كله يعتمد على عوامل من بينها اللغة المستهدفة وحجم وتنوع البيانات التي يقوم النظام بالتدرب عليها، بالإضافة إلى طبيعة البيئة التي سُجل فيها الصوت وغير ذلك [٢][٣].

تعتمد الطريقة التقليدية لمعالجة الصوت على تقسيم الجملة الصوتية إلى مجموعة ثابتة من الأطر fixed frame بحيث لا يزيد طول الإطار عن ٣٠ ميليثانية وذلك لثبات خصائص الكلام الصوتي في هذه الفترة وعدم تغير خصائصه. ولكن هذه الطريقة قد لا تلائم الوضع الطبيعي للكلام الصوتي حيث أن الصوت البشري يصدر على هيئة مقاطع صوتية segment units مختلفة الأطوال لا أطر زمنية frames [٥].

تُعرّف المقاطع الصوتية segment units على أنها الجزء الأساسي الأصغر في اللغة والمكونة من مقاطع ساكنة C (Consonants) ومقاطع متحركة V (Vowels). وفي اللغة االعربية يتم تقسيم وحدات الكلام إلى خمسة أنواع أساسية: حرف متحرك CV مثل (مَ)، حرف ممدود CVV مثل (ما) ، مقطع من متحرك فساكن CVC مثل (مَلْ) ، مقطع من ممدود فساكنCVVC مثل (مال)، ومتحرك فساكنين CVCC مثل (عَصْر). وبالتالي فإن كل مقطع صوتي Segment unit في اللغة ستكون عبارة عن صنف (class) وسيكون دور المصنف classifier التعرف على هذه المقاطع الصوتية من مجموعة كبيرة من عدد الأصناف classes الموجودة في اللغة في وقت واحد ، وهذا الأمر يعتبر صعبًا

من الناحية العملية خاصةً عندما يكون عدد الأصناف كبيرًا والتشابه بينهم أيضا كبيرًا [٦].

في نظام تلاوة القرآن الكريم، يبلغ عدد جميع المقاطع الصوتية segment units في الجزء الثلاثين من القرآن الكريم ٤٣٠٠ مقطعا صوتيا تقريبا، كما يبلغ إجمالي عدد أصناف هذه المقاطع الصوتية ٨٠٠ صنفا مختلفا تقريبا [٥]. وبالتالي يصعب تصنيف هذا العدد الكبير من الأصناف باستخدام الخوارزميات التقليدية؛ لذلك، فإننا نقترح في هذا البحث اتباع تقنية من تقنيات الذكاء الاصطناعي تسمى «التصنيف الشجري الهرمي» (Hierarchical Tree Classification). حيث يتم تجميع عدد كبير من الفئات في مجموعات فرعية قبل تصنيفها نهائيا [٢].

يشكل نظام التصنيف الهرمي هيكلًا يشبه الشجرة، حيث يمكن عبور العديد من المسارات من الجذر وصولًا إلى الأطراف (الأوراق) على مبدأ «فرق واغز» «Divide and Conquer»، حيث يتم تقسيم المشكلة الكبيرة بشكل متكرر إلى مشاكل أصغر وأسهل يمكن دمج حلولها لإيجاد حل للمشكلة الشاملة [٤][٥].

يتميز نظام التصنيف الهرمي عن المصنفات التقليدية بتقليل عدد الأصناف إلى أصناف أساسية والتي بدورها تقوم بالاستغناء عن الحسابات غير الضرورية. كما يُظهر التصنيف الهرمي مرونة في اختيار مجموعات فرعية مختلفة للفصول حسب قواعد للانتقال بين المراحل المختلفة في الشجرة، بالإضافة إلى إمكانية إجراء مفاضلة بين دقة التعرف على الصنف في أحد فروع الشجرة وكفاءة الفترة الزمنية للحصول على النتيجة.

أما عيوب تصنيف التسلسل الهرمي فمنها أن أي خطأ في نظام التعرف على أفرع الأشجار الرئيسية في المراحل الأولى تُوَرّث وتُنتَقل إلى المراحل الفرعية. وهنا تتجلى إشكالية المفاضلة بين الدقة والكفاءة، حيث يصعب تحسين كل من الدقة والكفاءة معا. علاوة على ذلك، فثمة صعوبات في تحديد القواعد وعدد المراحل في الشجرة الأمثل في التطبيق، وهذا بدوره يؤثر على نتيجة الأداء باستخدام التصنيف الهرمي [٦][٧].

في هذا البحث نقوم بعرض نظام مقترح لتصنيف الكلمات القرآنية باستخدام أساليب وتقنيات المصنفات التقليدية والمصنف الهرمي. سيتم تقسيم البحث إلى

ستة وحدات على النحو التالي: الوحدة الثانية عبارة عن وصف بنية نظام التعرف على الكلمات القرآنية. وفي الوحدة الثالثة يتم توضيح كيفية استخراج الخصائص للصوت (Feature Extraction) وفي الوحدة الرابعة نبين كيفية تقليل حجم مصفوفة الخصائص المستخرجة Feature Extraction matrix Dimension باستخدام تقنية تحليل المكونات الأساسية (Principle Component Analysis (PCA. وفي الوحدة الخامسة شرح أشهر المصنفات التقليدية والتي يتم استخدامها لمعالجة الصوت في القرآن الكريم. وفي الوحدة السادسة شرح نظام تصنيف التسلسل الهرمي وتطبيقه على الوحدات الكلامية في القرآن الكريم.

٢- بنية نظام التعرف الآلي على الوحدات الكلامية في القرآن الكريم

في هذا البحث سنقوم بالاعتماد على المقاطع الصوتية segment units في القرآن الكريم عوضا عن الإطارات الثابتة fixed frames. يوضح الشكل ١ الخطوات الرئيسية لبنية نظام التعرف الآلي وفيه الخطوات التالية:

١. الحصول على المقاطع الصوتية الخاصة بالقرآن الكريم.

٢. استخراج الخصائص المتعلقة بالمقاطع الصوتية القرآنية.

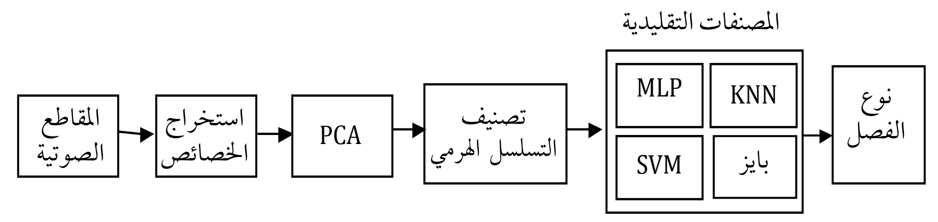
٣. تقليل أبعاد متجه الخصائص Feature Vector Dimension Reduction

٤. استخدام تقنية تصنيف التشجير الهرمي (Hierarchical Tree (HTC Classification لـتقليل عدد الأصناف إلى أصناف رئيسية.

٥. المقارنة مع خوارزمية المصنفات التقليدية وفي هذا البحث سنقوم بشرح أربع أنواع:

مصنف بايز Naïve Bayes [٢٠]، ومصنف الشبكة العصبية متعددة الطبقات (Multi-Layer Perceptron (MLP [٩]، ومصنف الجار الأقرب K-Nearest (Neighbor (KNN [١٠]، ومصنف آلة متجه الدعم Support Vector Machine (SVM) [١١].

وسنقوم الآن بشرح تفصيلي لكل خطوة من الخطوات الموجودة في شكل ١.



شكل (١): مخطط منهجية البحث في استخدام الذكاء الاصطناعي للتعرف على مقاطع القرآن الكريم

١,٢ الحصول على المقاطع الصوتية الخاصة بالقرآن الكريم

مدخلات النظام المقترح عبارة عن مقاطع صوتية خاصة بالقرآن الكريم حصلنا عليها من قاعدة بيانات مدينة الملك عبدالعزيز للعلوم والتقنية للجزء الثلاثين من القرآن الكريم [٥]، وبلغ إجمالي عدد وحدات المقاطع الصوتية فيها ما يقارب ٤٣٠٠ مقطعا صوتياً.

٢,٢ استخراج الخصائص المتعلقة بالمقاطع الصوتية القرآنية

استخراج الخصائص للمقاطع الصوتية مرحلة مهمة جدا في التعرف على الكلام. ويتمثل التحدي والصعوبة في كيفية استخراج خصائص قوية تمكن المصنف من التعرف على المقطع الصوتي وتحديد الصنف الذي ينتمي له هذا المقطع. ولاستخراج الخصائص من المقاطع الصوتية نقوم في البداية بتقسيم المقطع الصوتي المدخل إلى إطارات frames بطول نموذجي N، يتراوح من ٦٦٠ إلى ١٣٢٠ عينة لكل إطار، والتي تقدر من ١٥ إلى ٣٠ مللي ثانية، والتي تحافظ على ثبات خاصية الصوت في هذه الفترة الزمنية. لقد قامت دراسات سابقة كثيرة لتحديد الخوارزميات التي تقوم باستخراج الخصائص من الصوت وفي هذا البحث سيتم التطرق إلى أهم هذه الخصائص.

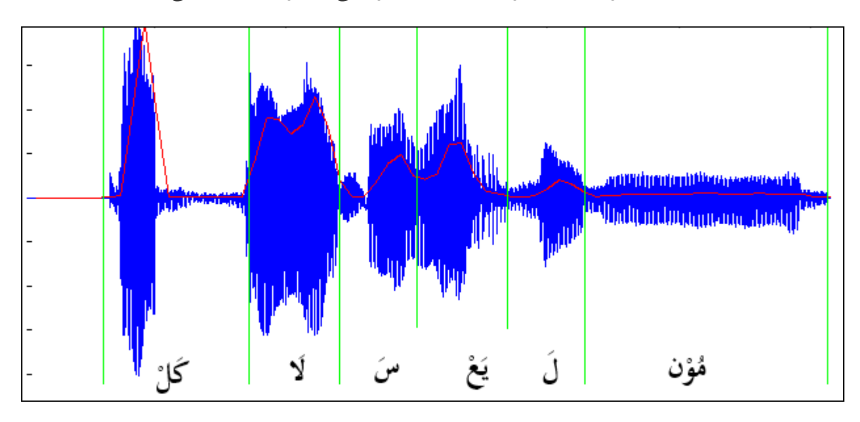
١,١,٢ خاصية الطاقة الصوتية

تعتبر خاصية الطاقة الصوتية سمة ممتازة خاصة للتمييز بين المقاطع الساكنة (consonant) والمقاطع المتحركة (vowels)، نظرًا لأنها تكون عادة ذات قيمة عالية في المقاطع المتحركة وقيم منخفضة في لمقاطع الساكنة. ولاستخراج خاصية الطاقة من المقطع الصوتي نقوم في البداية بتحويل المقطع الصوتي من مستمر Continuous إلى

متقطع Discrete عن طريق تقطيع المقطع الصوتي إلى عينات samples بفرق زمني ثابت ومن ثم يتم تطبيق المعادلة التالية [١٣]:

$$E_i \sum_{n=1}^{N} x_i(n)^2 \quad \ldots\ldots (١)$$

حيث تمثل E_i الطاقة الكلية للمقطع الصوتي i و تمثل $x_i(n)$ عينة n (sample) في المقطع الصوتي i و N هو العدد الكلي للعينات (samples) في المقطع الصوتي. ومثال على ذلك فإن قيمة الطاقة في الآية ﴿كلا سيعلمون﴾ تظهر بلون أحمر في الشكل ٢. حيث نلاحظ أن قيمة الحرف المتحرك (كَ) أكثر من الحرف الساكن (لْ).



الشكل (٢): قيمة الطاقة للمقاطع الصوتية في آية ﴿كلا سيعلمون﴾ [٥]

٢,١,٢ خاصية حدة الصوت (Pitch)

تُعرف «حدة الصوت» على أنها خاصية إدراكية تسمح بترتيب الأصوات حسب سلم مرتبط بالتردد، أي حسب عدد تكرار الاهتزازات (الذبذبات) هيرتز في الثانية للطبقات الصوتية أثناء التحدث [١٤]. حيث يتم استخدام هذه الخاصية لمعرفة التردد الأساسي للمقطع الصوتي بناءً على الارتفاع والانخفاض في نغمة الصوت.

هناك طرق مختلفة يمكن استخدامها لتقدير درجة الصوت من إشارة الكلام . سنشرح فيما يلي طريقة «تقنية الارتباط التلقائي» Autocorrelation Technique بين كل إطار والإطار الآخر من خلال استخدام المعادلة التالية:

$$R(k) = \sum_{i=1}^{f_L} x(m)x(m+k) \quad \ldots\ldots (٢)$$

حيث أن FL هو طول الإطار، $x(m)$ هو إطار الإشارة، k عامل الإزاحة، و $R(k)$ هي دالة الارتباط التقريبي التلقائي.

٢,١,٣ خاصية ترددات صفة صوت الكلام Formant Frequencies

تُعرَّف خاصية ترددات صفة صوت الكلام على أنها ترددات الرنين والاهتزاز في الأحبال الصوتية أثناء النطق وتكون ظاهرة بشكل كبير في الحروف المجهورة (حروف كلمة قطب جد) أكثر من الحروف المهموسة (مثل حرف الحاء والهاء) [١٥]. ويمكن تمثيل هذه الترددات عن طريق حساب القيم العظمى للترددات Peaks of The Frequency Response من خلال تقنية الترميز التوقعي الخطي Linear Predictive Code (LPC) والتي تمثل على النحو التالي: [١٧]

$$\tilde{x}[n] = \sum_{k=1}^{p} a_k x[n-k] \quad \text{.........(٣)}$$

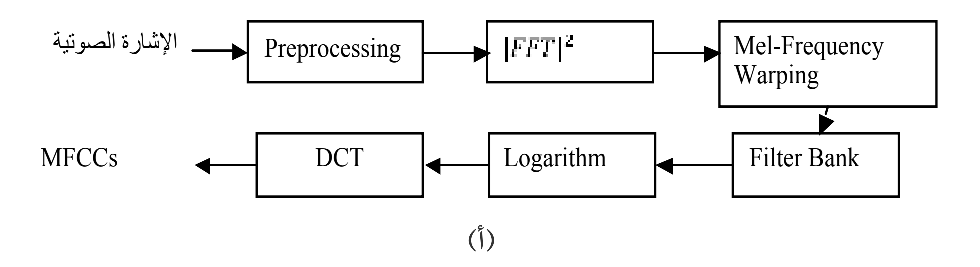
حيث $\tilde{x}(n)$ هي العينة المتوقعة عند الوقت n، والمتغير p عبارة عن عدد العينات السابقة للوقت n، و a_k هي معاملات LPC .

٢,١,٤ خصائص معاملات تردد ميل Mel-Frequency Cepstrum

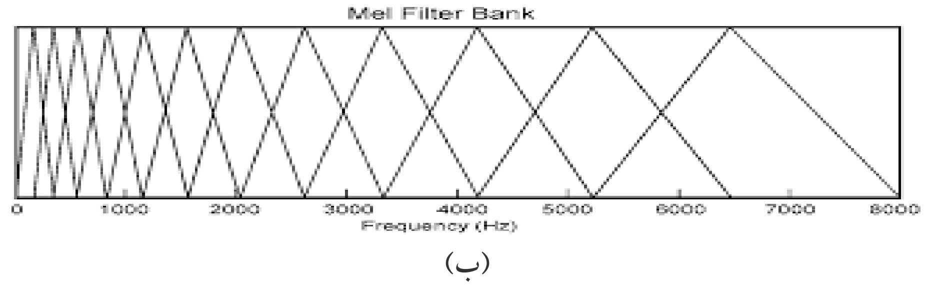
تعتبر تقنية معاملات تردد ميل (MFCCs) من أكثر الخصائص استخداما للتعرف على الكلام. حيث أن الفكرة وراء معالجة MFCC هي مقاربة الطريقة التي يسمع بها البشر الأصوات. حيثركز الأذن البشرية عند الاستماع على الترددات المنخفضة، وهذا ما تحاوله MFCC من خلال تكبير مدى هذه الترددات باستخدام اللوغاريتمات. يبدأ استخراج MFCC لكل إطار في المقطع الصوتي والذي يتراوح من ٦٦٠ إلى ١٣٢٠ عينة لكل إطار، والتي تقدر من ١٥ إلى ٣٠ مللي ثانية [١٤]. لتحويل الترددات الخطية إلى مقياس ميل تكون في المعادلة التالية:

$$M(f) = 1125\, ln(1 + \frac{f}{700}) \quad \text{.........(٤)}$$

حيث f قيمة التردد في هرتز. لتوضيح خطوات عمل MFCC موضحة في الشكل ٣.



(أ)

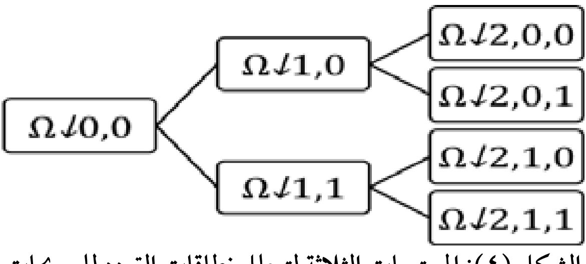


(ب)

الشكل (٣): (أ) خوارزمية MFCC (ب) مرشح ميل

٥ , ١ , ٢ تحويل المويجات المنفصلة Discrete Wavelet Transform (DWT)

يعتبر تحويل المويجات المنفصلة من الخصائص المميزة في تمثيل الإشارة في كل من مجال الزمن والتردد Time and Frequency domain، حيث هذا المجالان هما التمثيلان المشهوران للإشارات، حيث يُبرز كل منهما جانبا من خصائص الإشارة. إن الفكرة الرئيسية من DWT هو تقسيم نطاق إشارة تردد المقطع الصوتي وترتيبها من الأقل إلى الإعلى بشكل متعاقب كما هو موضح في الشكل ٤ . حيث $\Omega_{0,0}$ (العقدة الجذرية لشجرة نطاق الترددات في المقطع الصوتي) تمثل تردد الإشارة الأصلية. ومن ثم يتم تقسيم نطاق التردد إلى قسمين بحيث $\Omega_{1,0}$ تمثل النصف العلوي من نطاق تردد المقطع الصوتي و $\Omega_{1,1}$ تمثل الجزء السفلي من مجال تردد المقطع الصوتي وهكذا حسب عدد المستويات المطلوبة.



الشكل (٤): المستويات الثلاثة لتحلل نطاقات التردد للمويجات

٣,٢ تقليل أبعاد متجه الخصائص Feature Vector Dimension Reduction

إن استخـــدام الخصائص التي تـــم الحديث عنهـــا في الفقرة السابقة يعـــطي نتائج جيدة ولكن ليست ممتازة وذلك بسبب احتمالية وجود بيانات كثيرة مكررة أو ليست ذات أهمية في التمييز بين الأصنـــاف مما قد يؤدي إلى تعقد اعملية التصنيف. ولكي نقوم بتحسين هذه النتائج يتم استخدام تقنية تحليل المكونات الأساسية Principle Component Analysis (PCA) لتقليص بيانات الخصائص واختصارها، حيث تقوم بتحويل العدد الكبير من المتغيرات المترابطة ضمنا –ولو بشكل جزئي– إلى مجموعة أصغر من المتحولات المستقلة التخيلية، وهي تدعى عادة بالمكونات الرئيسية وتحسب أساسا من المتغيرات الأصلية بنسب ومقادير تزيد أو تنقص بحسب دور وتأثير كل منها، لتصف أكبر قدر ممكن من البيانات الموجودة في خصائص الأصناف.

إن الفكرة الأساسية في تحليل المكونات الرئيسية PCA هو تقليل حجم مصفوفة استخراج الخصائص إلى أكبر قدر ممكن والتي تسهم في التمييز بين الأصناف، وذلك من خلال عمل محاور تخيلية متعامدة والتي تحسب من خلال مجموع الخصائص المستخرجة للمقاطع الصوتية الحقيقية لكن بأوزان متفاوتة تعكس دور كل منها وأهميته في التفريق ما بين الأصناف. تعمل خطوات تنفيذ الخوارزمية على حصر أكبر قدر ممكن من التباينات ضمن توليفة الخاصية التخيلية الأولى والتي عادة ما يطلق عليها تسمية المكون الأساسي الأول PC1، كما يتم حساب نسبة مؤوية لهذه الخاصية التخيلية والتي تشير إلى الحصّة الكليّة من التباينات التي تم إلتقاطها والتعبير عنها في هذه الخاصية التخيلية. ثم بعد ذلك يأتي الدور في تكوين المكون الأساسي الثاني PC2 والذي سيقوم بدوره بمحاولة التعبير عن أكبر قدر ممكن من التباينات المتبقية والتي لم يستطع PC1 التعبير عنها، ويستمر الأمر بالنسبة لكل من PC3 وPC4 وصولا إلى العدد الكلي للخصائص التي تم استخراجها للمقاطع الصوتية.

بهذه التقنية نستطيع التمييز بين الخصائص التي لا تسهم في التفريق ما بين الأصناف المختلفة في مجموعة البيانات ويكون لها أوزان صغيرة تقترب من الصفر، وبين الخصائص التي لها دورا هاما في التفريق ما بين الأصناف حيث يكون لتلك الصفات أوزان ذات مقادير كبيرة تقتربفي قيمتها المطلقة من الواحد الصحيح [٢١].

ولحساب PCA من الناحية الرياضية نقوم في البداية بتحليل القيمة الذاتية eigenvalues لمصفوفة التباين التقريبي estimated covariance. وهذا الأمر يتم من خلال إيجاد الوسط الحسابي لمصفوفة البيانات الخاصة بكل نوع من أنواع المقاطع الصوتية. ويمكن إيجاد مصفوفة التباين التقريبي من خلال العلاقة التالية:

$$S_X = \frac{1}{n} X X^T \qquad \text{.....(٥)}$$

حيث X هي مصفوفة الخصائص المستخرجة من جميع المقاطع الصوتية في قاعدة البيانات والتي أبعادها $n \times m$ حيث أن m هي عدد الخصائص الكلية التي تم استخدامها، و n هو عدد الملاحظات observations والتي تعني هنا جميع المقاطع الصوتية، والمتغير S_X عبارة عن مصفوفة مربعة متماثلة أبعادها m x m.بحيث أن قطر المصفوفة S_X عبارة عن قيم التباينات التقديرية بين المتغيرات. للحصول على تحويل PCA نقوم بتطبيق المعادلة التالية:

$$Y = PX \qquad \text{.....(٦)}$$

حيث Y عبارة عن تمثيل X بناء على أساس المصفوفة الجديدة P، حيث أن P عبارة عن مصفوفة تحول X إلى نظام الإحداثيات التخيلية الجديدة وتكون فيها البيانات مرتبة من الأكبر إلى الأصغر. ولإيجاد مصفوفة تقدير التباينات بالنسبة للمصفوفة Y يتم احتسابها من خلال المعادلات التالية:

$$\begin{aligned} S_Y &= \frac{1}{n} Y Y^T \\ &= \frac{1}{n} (PX)(PX)^T \\ &= \frac{1}{n} P X X^T P^T \\ &= P(\frac{1}{n} X X^T) P^T \end{aligned}$$

$$S_Y = P S_X P^T \qquad \text{.....(٧)}$$

كما أن مصفوفة تقدير التباينات S_X يمكن تحليلها باستخدام تحليل القيم الذاتية على النحو التالي:

$$S_X = UD \quad \text{...(٨)}$$

حيث أن D عبارة عن مصفوفة قطرية تكون فيها البيانات مرتبة حسب القيم الذاتية من الأكبر إلى الأصغر. والمصفوفة U عبارة عن المتجهات الذاتية eigenvectors حيث أن كل عمود في المصفوفة عبارة متجه ذاتي والتي تتميز بأنه عندما يتم إجراء تحويل خطي على هذه المتجهات لا يتغير اتجاهها. وبما أن المصفوفة S_X متماثلة فإن $U^T=U^{-1}$ وبالتالي يمكن كتابة S_X على الشكل التالي:

$$S_X = UDU^T \quad \text{...(٩)}$$

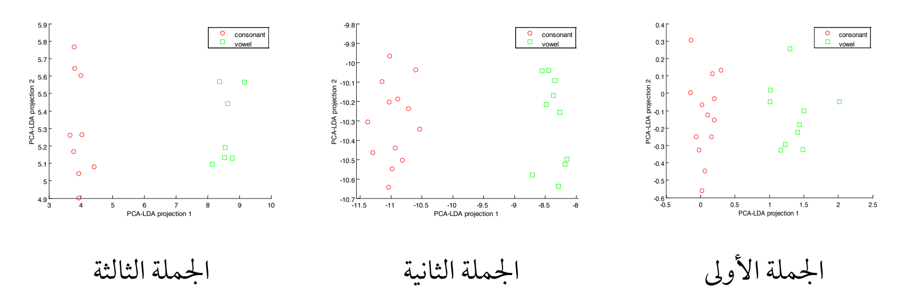
وبالعودة إلى S_y، نفترض أن $P = U^T$ فإن S_y تكون على الشكل التالي:

$$S_Y = PS_XP^T \quad = U^TS_XU \rightarrow \quad = U^T(UDU^T)\,U$$

$$= (PP^T)D(PP^T) \rightarrow \quad S_Y = D \rightarrow (10)$$

يمكننا أن نرى أنه عندما يتم اختيار مصفوفة التحول على أساس $P = U^T$ ، $P = U^T$ فإن ناتج الخصائص المتحولة (العناصر الموجودة في المصفوفة Y) تصبح غير مهمة بما أن مصفوفة التغاير في النتائج قطرية. إن هذه الطريقة أدت إلى عمل ترتيب القيم الذاتية والمتجهات الذاتية حسب الأهمية وبالتالي يمكن تقليل أبعاد المصفوفة إلى d x n حيث أن d عبارة عدد الصفوف المطلوبة من المصفوفة الكلية. وعادة في الأبحاث يكون اختيار حجم المصفوفة بحيث يكون مجموع التباينات على الأقل ٨٠٪ من مجموع التباينات الكلية.

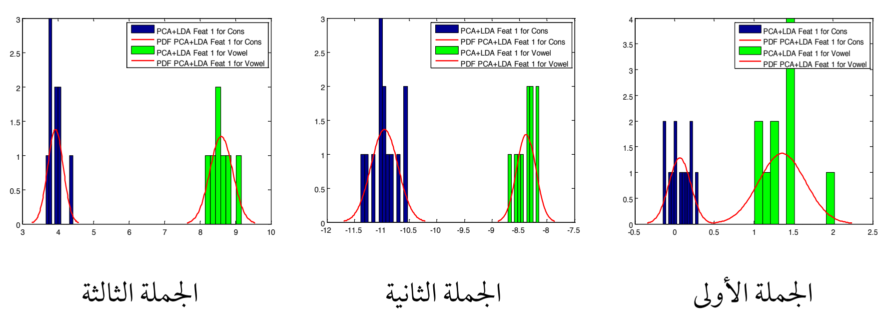
ولتوضيح أهمية وقوة PCA، نوضح في الشكل ٥ مخطط التشتت scatter plot (مخطط يستخدم بيانياً لتقديم وعرض العلاقة بين متغيرين) لكل من الحروف الساكنة consonants والحروف المتحركة vowels للمقاطع الصوتية بعد تطبيق نظام التحويل PCA.



الجملة الثالثة الجملة الثانية الجملة الأولى

الشكل (٥): مخطط التشتت للحروف الساكنة والمتحركة بعد تقليل أبعاد الخصائص باستخدام تحويل PCA.

في الشكل ٦، يوضح الرسم البياني لكل من الحروف الساكنة والحروف المتحركة بناء على دالة التوزيع الاحتمالي PDF على شكل توزيع جاوس Gaussian distribution. حيث نلاحظ أيضا أن صنف الحروف الساكنة منفصلة تماما عن صنف الحروف المتحركة وهذا بدوره يؤدي إلى الحصول على نتائج ممتازة للتصنيف بين الأصناف.



الجملة الثالثة الجملة الثانية الجملة الأولى

الشكل (٥): الرسم البياني لدالة توزيع الاحتمالات على شكل جاوس بعد تطبيق إسقاط PCA.

٤, ٢ التصنيف الهرمي Hierarchical Classification

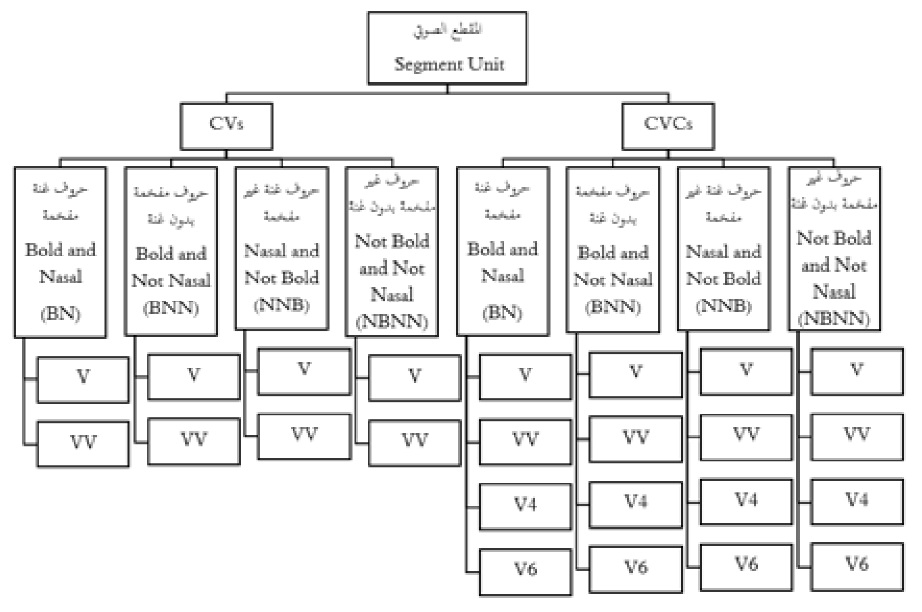
يعتبر تصميم هيكل شجرة التصنيف الهرمي (Hierarchical Classification أو HTC) من الطرق المهمة في التصنيف وذلك من خلال البحث عن الشجرة المناسبة والخصائص المناسبة للمجموعات الفرعية حتى يتم التعرف على الفروع في كل طبقة . إن أبسط طريقة هي تقسيم المشكلة إلى مشكلات فرعية لا تحتوي على عناصر مشتركة، وتسمى أيضًا «الانقسام الصعب» [١٩] .ويمكن استخدام هذه الطريقة في القرآن الكريم بحيث يتم تصنيف المقاطع الصوتية على شكل هرمي HTC كما هو موضح في الشكل ٧.

تعتمد بنية HTC على المعرفة المسبقة كيفية قراءة المقاطع الصوتية بناءً على قواعد التجويد المستخدمة في تلاوة القرآن الكريم . في بداية التصنيف الهرمي في الطبقة الأولى يتم التمييز في جذر الشجرة بين المقاطع الصوتية من نوع CV أو CVC وذلك عن طريق أخذ آخر ثلاث إطارات frames من المقطع الصوتي ونقوم باستخدام خاصية الطاقة energy للتعرف هل نهاية المقطع هل هو حرف ساكن أم متحرك.

في الطبقة الثانية يتم تصنيف كل فرع بناء على معيارين رئيسين: المعيار الأول هل الحرف الساكن مفخم أم لا، والمعيار الثاني هل الحرف الساكن فيه غنة أم لا. بناء على هاذين المعيارين فلقد تم تجزئة الفرع الأول من الشجرة CV إلى أربعة أجزاء: الجزء الأول حرف ساكن مفخم بغنة (مثل كلمة «قُتل»)، والجزء الثاني ساكن مفخم بدون غنة (مثال ذلك كلمة «طُبع»)، والجزء الثالث ساكن غير مفخم بغنة (مثال ذلك كلمة «كنتم»)، والجزء الرابع ساكن غير مفخم بدون غنة (مثال ذلك كلمة «سَأل»). أما في الطبقة الثالثة في هذا الفرع فكان المعيار الرئيسي كم زمن الحرف المتحرك، حيث في القرآن الكريم يكون إما حركة أو حركتين أو أربع أو ست حركات بناء على قواعد التجويد. بناء على معيار زمن الحرف المتحرك فسيكون إما حركة واحدة (V) (مثال ذلك الفتحةَ)، أو حركتين (V2) (مثال ذلك المد بالألف) . وبنفس هذه المعايير في الفرع الأول من الشجرة الرئيسية قمنا بتطبيقها على الفرع الثاني من الشجرة الرئيسية .CVC حيث تم تقسيم CVC كذلك إلى حرف ساكن مفخم بغنة (مثال ذلك «من قال»)، وحرف ساكن مفخم بدون غنة (مثال ذلك كلمة «قال»)، وحرف ساكن غير مفخم بغنة (مثال على ذلك كلمة «أنتم»)، وحرف ساكن غير مفخم بدون غنة (مثال ذلك كلمة «قيل») . ثم، ينقسم كل فرع إلى أربع أجزاء النوع الأول متحرك قصير (V) (مثال على ذلك)، حرف ممدود (V2) (على سبيل المثال ا)، حرف ممدود بزمن أربع حركات (V4) (على سبيل المثال سائل)، وحرف ممدود بزمن ست حركات (V٦) (مثال على ذلك سيعلمونعند الوقوف عليها يكون مد عارض للسكون بمقدار ٦ حركات) . نلاحظ في الشكل ٧ أن الفرع V4 وV6 ليست مدرجة في فرع CV. حيث هذا النوع V٤، يحدث عندما يتبع الحرف المتحرك حرف همزة (ء) وهذا لا يكون إلا إذا كان المقطع من نوع CVC. كذلك الفرع من نوع V6 يحدث عندما يكون بعد الحرف المتحرك حرفًا ساكنا عندما يتوقف القارئ عن قراءة الآية. بناء على هذه الأنواع يكون

لكل مقطع صوتي نوع واحد فقط من هذا الأفرع وبالتالي يسهل عملية التصنيف.

بناء على ما تم شرحه في التصنيف الهرمي، فمن الواضح بأن HTC لها ثلاث طبقات: الطبقة الأولى لدينا فئتين رئيسيتين CVs و CVCs. في الطبقة الثانية لديناأربع تصنيفات تحت كل فرع: مفخم بغنة، مفخم بدون غنة، غير مفخم بغنة، و غير مفخم بدون غنة. أما في الطبقة الثالثة، لدينا الفئات الفرعية V و V2 تحت فرع مقاطع CV والفئات الفرعية V وV2 وV4 وV6 تحت فرع CVCs. وبالتالي يبلغ إجمالي عدد التفريعات ٢٢ تفريعة أي أنه تم تقليص عدد الأصناف للمقاطع الصوتية من ٨٠٠ إلى ٢٢ صنفا.



الشكل ٧: شجرة تصنيف المقاطع الصوتية الخاصة بالقرآن الكريم

٣- خوارزميات التصنيف Classification

تهدف عمليات التصنيف (ضمن بيئة تعلم الآلة المراقب (Supervised Learning)) لتصنيف بيانات التدريب ضمن فئات مختلفة حسب خواصها المشتركة ولها عدة خوارزميات . هذا، وتعتمد عملية التصنيف على النماذج (Models) التي يتم بناؤها أثناء عملية التصنيف والمرتبطة بنوع المصنف (Classifier) المستخدم [٢٠].

وفيما يلي نستعرض بعض المصنفات التي تمت المقارنة معها في هذا البحث:

١ , ٣ مصنف بايز Naïve Bayes

يستند هذا المصنف إلى نظرية بايز الاحتمالية (Bayes' theorem) القائمة على مبدأ الاحتمال الشرطي الذي يقوم بحساب احتمال وقوع أحد الأحداث الاحتمالية بناء على وقوع حدث مستقل آخر أو أكثر وفق المعادلة التالية:

Prob(B given A) = Prob(A and B) / Prob(A) (١١)

حيث:

Prob(B given A): احتمال وقوع الحدث B بناء على وقوع الحدث A – وهو الاحتمال المطلوب

وProb(A and B): احتمال وقوع الحدثين A و B معاً أو ما يدعى (pairwise)

وProb(A): احتمال وقوع الحدث Aأو ما يدعى (singleton).

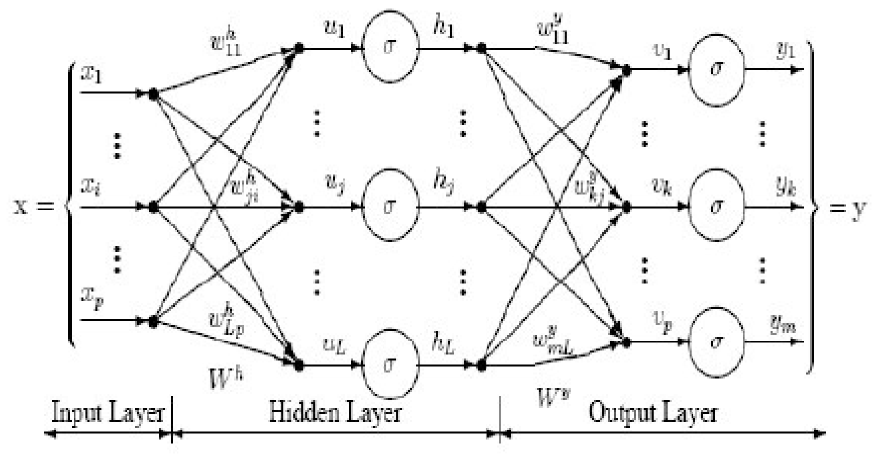
إن الحدث (B) يبدو كحدث مرتبط بحدث مستقل وحيد (A)، لكن في الحقيقة تقوم هذه الخوارزمية أغلب الأحيان بربط الحدث بعدة أحداث مستقلة.

يمتاز هذا التصنيف بالسرعة في بناء النماذج كما أنه يمتاز بأنه قابل للتوسع (scalable) مع ازدياد بيانات التدريب وبتنفيذ عملية بناء النماذج بشكل متوازي (parallelized) ويمكن استخدامه لتصنيف بيانات ثنائية الفئات (binary class) أو متعددة الفئات (multi class).

٢ , ٣ مصنف الشبكة العصبية متعددة الطبقات(Multi-Layer Perceptron (MLP

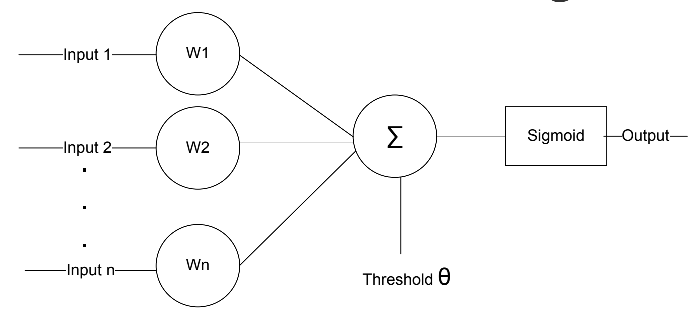
يعتمد هذا المصنف على خوارزميات الشبكة العصبية (Artificial Neural Network) ذات طبقة أو أكثر بين الدخل والخرج بحيث ترتبط كل عقدة (Node) في كل طبقة بجميع العقد الأخرى في باقي الطبقات، وباستثناء طبقة الدخل فإن جميع العقد هي عصبونات اصطناعية (Artificial Neuron)،كما هو موضح في الشكل ٨.

عند تدريب البيانات أو اجراء الاختبار عليها يتم إدخال البيانات عبر طبقة الإدخال (Input Layer) وتتم معالجتها ضمن الطبقات المخفية (Hidden Layers) وعرضها بالنهاية عبر طبقات الخرج (Output Layer).



الشكل (٨): أنواع الطبقات الثلاث لتصنيف MLP.

تتألف كل طبقة من واحدة أو أكثر من العصبونات الاصطناعية المتوازية، لكل عصبون كما يظهر في الشكل ٩ عدد N من المدخلات ذات الوزن W لكل منها بالإضافة لمخرج واحد فقط. يقوم كل عصبون بدمج المدخلات مختلفة الأوزان من خلال جمعهم سوية وبالاستناد إلى حد العتبة Threshold والذي يرمز له عادة بالحرف الإغريقي θ ليقوم بتحديد قيمة الخرج.



الشكل (٩): بنية العصبون الاصطناعي

لشرح آلية عمل هذه الخوارزمية بصورة مبسطة لابد من تعريف المتغيرات التالية:

المدخلات (x_1، x_2، ....، xn) ذات الأوزان (w_1، w_2، .... wn).

الدالة u دالة تعبر عن احتمالية التنشيط (activation potential).

دالة حد العتبة θ (threshold).

دالة الخرج y (output).

دالة التنشيط f (activation function)

يعرف دالة احتمالية التنشيط بالمعادلة:

$$u = \sum_{i=1}^{N}(w_i x_i) \quad (12)$$

وبالاعتماد على تعريف دالة الخرج المبينة في المعادلة:

$$y = f(u - \theta) \quad (13)$$

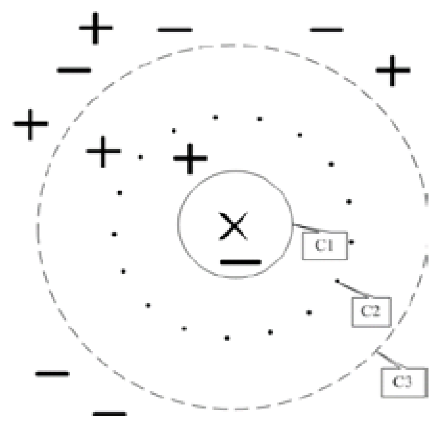
المعادلة النهائية لتابع الخرج تظهر في المعادلة:

$$y = f\left(\sum_{i=1}^{N}(w_i x_i)\right) : w_0 = \theta،\ x_0 = -1 \quad (14)$$

يتم استخدام هذا التصنيف بشكل واسع في عدة مجالات؛ كالتعرف الآلي على الكلام (speech recognition)، والتعرف الآلي على الصور (image recognition) إضافة لبرامج الترجمة الآلية (machine translation).

٣,٣ مصنف الجار الأقرب K-Nearest Neighbor

مصنف الجار الأقرب (K-Nearest Neighbor أو KNN) تهدف للتنبؤ بالصنف عن طريق مقارنة السجلات الشبيهة بالسجل المراد التنبؤ بقيمته وتقدير القيمة المجهولة لهذا السجل بناء على مقدار تلك السجلات. يعتمد عمل هذه الخوارزمية بشكل أساسي على وحدة القياس (metric). يمثل الرمز (K) عدد الحالات الأكثر تشابهاً مع الحالة المراد التنبؤ بقيمتها. الشكل (١٠) يوضح آلية عمل هذه الخوارزمية حيث تظهر النقطة المجاورة الأقرب لإحدى نقاط البيانات المراد تصنيفها (X) ضمن الحد الفاصل (المسافة) (C1) بينما يظهر ضمن الحد الفاصل (C2) النقطتين المجاورتين للنقطة (X) وضمن الحد الفاصل (C3) النقاط الثلاثة المجاورة للنقطة (X).



الشكل (١٠): توزع البيانات ضمن المصنف KNN.

تنتمي النقطة (X) في حالة (C1) تنتمي إلى الصف السالب، وفي حالة (C3) إلى الصف الموجب وذلك حسب نظام التصويت للأغلبية (Majority Voting Scheme)، أما في حالة (C3) فإنه يتم اختيار الصف بناء على وحدة القياس (metric) ليتم تصنيف النقطة على أساسه. يتم اختيار العدد (K) بشكل مناسب مع عدد البيانات بحيث يتم التغلب على التراكب الناتج عن عملية التصنيف والتي تزداد مع ازدياد شذوذ البيانات وعدم تناسقها.

٤,٣ مصنف آلة متجه الدعم Support Vector Machine (SVM)

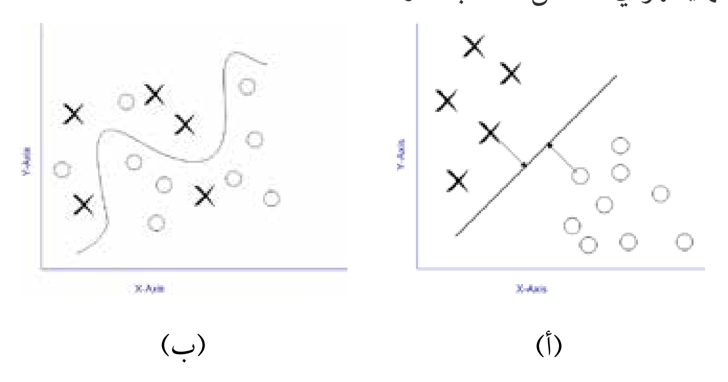
يعتبر هذا المصنف أحد أقوى المصنفات التقليدية بامتلاكه آلية عمل تدمج كلاً من خوارزمية الشبكات العصبونية مع خوارزمية الشعاع الأساسي (Radial Basis) لإيجاد أفضل سطح فاصل بين بيانات التدريب. يمتاز هذا المصنف بالمرونة، قابلية التوسع والسرعة في الأداء مما يعطيه الأفضلية في معالجة مسائل التعرف الآلي المتنوعة وعلوم معلوماتية الأحياء (Bioinformatics)، كما يتميز هذا المصنف بقدرته على معالجة معطيات ذات عدد كبير من المعايير مقارنة بعدد سجلات البيانات المتواجدة.

تعمل آلية تصنيف المعطيات وفق حالتين:

- تصنيف خطي: وذلك باختيار أفضل خط مستقيم أو مستوي يستطيع فصل البيانات ويكون أقرب ما يمكن لجميع هذه البيانات وهنا يمكن تمييز حالتين:

مجموعة البيانات ثنائية الفئة (Binary Class) ذات بعدين (D=2)، يبين الشكل (١١) مستقيم الحالة الأمثل (Optimum Situation) التي تقسم مجموعة البيانات إلى قسمين ويمكن تعميم هذه الحالة إلى مجموعة البيانات متعددة الفئات (Multi Class) متعددة الأبعاد (D>2)،

- تصنيف غير خطي: وذلك باختيار أفضل سطح أو منحني يستطيع فصل البيانات ويكون أقرب ما يمكن لجميع هذه البيانات وهنا يمكن تميز حالتين حسب فئات وأبعاد البيانات فيها إذا كانت مؤلفة من ثنائية الفئة فقط (D=2) كما يظهر في الشكل (٥-ب) أو متعددة الفئات (D>2).



الشكل(١١): توضيح أسطح فصل البيانات لمصنف SVM.

٤- التجارب والنتائج

بناء على ما تم ذكره في الفقرات السابقة، سنقوم في هذه الوحدة بتطبيق هذه المفاهيم من الناحية العملية وعرض نتائج البحث. كما قلنا سابقا بأنه تم الاعتماد على قاعدة بيانات مدينة الملك عبد العزيز للعلوم والتقنية والحصول المتكونة من ٤٣٠٠ مقطعا صوتيا [٥]. حيث أن هذه المقاطع الصوتية تم تصنيفها على صيغة CV وصيغة CVC بناء على مبدأ التصنيف الشجري الهرمي المبني على قواعد التجويد والتي تحتوي على ٢٢ صنفا رئيسيا. في بداية الأمر تم استخراج ٢٨١ خاصية لكل مقطع صوتي على النحو التالي:

- خوارزمية الطاقة وتم استخراج خاصية الطاقة لكل مقطع صوتي.
- خوارزمية درجة حدة الصوت وتم استخراج أربع خصائص وهي معدل وتشتت وأعلى وأقل قيمة درجة حدة صوت المقطع الصوتي.
- خوارزمية ترددات صفة صوت الكلام وتم استخراج ثلاث خصائص وهي معدل وتشتت وأعلى قيمة ترددات صفة صوت الكلام للمقطع الصوتي.
- خوارزمية معاملات تردد ميل MFCC وتم استخراج عشرين خاصية عن طريق إيجاد المعدل والتشتت لأول عشر معاملات الخوارزمية للمقطع الصوتي.
- خوارزمية تحويل المويجات المنفصلة للطبقات السبعة حيث تم استخراج ٢٥٥ خاصية للمقطع الصوتي.

بعد استخراج هذه الخصائص للمقاطع الصوتية أصبح حجم مصفوفة استخراج الخصائص ٤٣٠٠x٢٨١ عنصرا. ثم بعد ذلك تم استخدام تقنية تحليل المكونات الأساسية PCA بحيث تم تقليل حجم المصفوفة إلى ٤٣٠٠x٥٠ والتي تحتوي على مجموع نسبة التشتت ما يقارب ٩٠٪ من نسبة التشتت للخصائص الحقيقية. ثم بعد ذلك تم إدخال مصفوفة البيانات ٤٣٠٠x٥٠ إلى نظام تصنيف التشجير الهرمي المبني على قواعد تجويد القرآن الكريم لتصنيفها إلى CV وCVC كما تم توضيحه سابقا. ثم بعد ذلك تم استخدام المصنفات التقليدية (MLP. KNN. SVM. NB) حيث تم تدريب هذه المصنفات على ٨٠٪ من البيانات وعمل فحص ٢٠٪ المتبقية من البيانات.

أعطى المصنف SVM أفضل النتائج حيث كانت نتيجة دقة البيانات ما يقارب ٨٦٪ للمقاطع الصوتية من نوع CV و٩٠٪ للمقاطع الصوتية من نوع CVC.

إن هذه النتيجة لو قارنها بدون استخدام التصنيف الهرمي لحصلنا على نتيجة ٤٩٪. مما يعني أنه باستخدام التصنيف الشجري المبني على قواعد التجديد يتم تحسين النتائج بنسبة ٣٤٪.

٥- الخاتمة

في هذا البحث تم عمل دراسة عن التعرف الآلي على الكلام العربي المنطوق وتطبيقاته في القرآن الكريم باستخدام أنظمة الذكاء الاصطناعي. حيث تم في البداية الحصول على المقاطع الصوتية القرآنية من خلال قاعدة بيانات مدينة الملك عبد العزيز ومن ثم تم استخراج الخصائص لهذه المقاطع الصوتية باستخدام خوارزميات مشهورة في مجال معالجة الصوت. تبين أن حجم مصفوفة استخراج الخصائص لهذه المقاطع الصوتية كبيرة وبالتالي تم استخدام تقنية تحليل المكونات الأساسية PCA لتقليل حجم المصفوفة واستخدام خصائص تخيلية تقوم بإعطاء الأوزان الأعلى للخصائص الحقيقية الأهم وأوزان قليلة للخصائص الحقيقية الغير مهمة والتي بدورها أسهمت بشكل كبير في تحسين النتائج. ثم بعد ذلك تم استخدام خاصية التصنيف الهرمي بناء على قواعد التجويد القرآنية والتي بدورها قللت عدد الأصناف من ٨٠٠ صنف إلى ٢٢ صنف. وفي النهاية تم عرض أشهر المصنفات التي تسخدم في معالجة الصوت بشكل عام وفي القرآن بشكل خاص.

المراجع

[1] S. J. Russell and P. Norvig. Artificial Intelligence. A Modern Approach. 2010.

[2] X. He and L. Deng. "Discriminative learning for speech recognition: Theory and practice." vol. 4. 2008.

[3] M. K. Sharma. "Speech Recognition : A Review." in Special Conference Issue: National Conference on Cloud Computing & Big Data. 2015.

[4] R. K. Aggarwal and M. Dave. "Implementing a Speech Recognition System Interface for Indian Languages." Proc. IJCNLP-08 Work. NLP Less Privil. Lang.. no. January. pp. 105–112. 2008.

[5] A. H. Abo. M. Deriche. M. Elshafie. Y. Elhadj. and B. Juang. "Algorithm for Arabic Speech using Feature Fusion and a Genetic Algorithm." IEEE Access. 2018.

[6] P. A. A. Ali and I. T. Hwaidy. "Hierarchical Arabic Phoneme Recognition Using Mfcc Analysis." Iraq J. Electr. Electron. Eng.. vol. 3. no. 1. 2007.

[7] R. Polikar. "Ensemble based systems in decision making." Circuits Syst. Mag. IEEE. vol. 6. no. 3. pp. 21–45. 2006.

[8] and M. A. Yahya Ould Mohamed Elhadj. Mansour Alghamdi. "Phoneme-Based Recognizer to Assist Reading the Holy Quran." Adv. Intell. Syst. Comput.. vol. 235. pp. 141–152. 2014.

[9] E. M. Essa. A. S. Tolba. and S. Elmougy. "A comparison of combined classifier architectures for arabic speech recognition." 2008 Int. Conf. Comput. Eng. Syst. ICCES 2008. pp. 149–153. 2008.

[10] N. N. Radio. "Neural Networks used for speech recognition." in NINETEENTH NATIONAL RADIO SCIENCE CONFERENCE. ALEXANDRIA. 2002. vol. 2. no. 4. pp. 19–21.

[11] J. Hai and E. M. Joo. "Improved linear predictive coding method for speech recognition." Information. Commun. Signal Process. 2003 Fourth Pacific Rim Conf. Multimedia. Proc. 2003 Jt. Conf. Fourth Int. Conf.. vol. 3. no. December. pp. 1614–1618 vol.3. 2003.

[12] F. O. F. Engineering. "Parametric Speech Emotion Recognition Using Neural Network." 2014.

[13] A. Lilia and R. Herrera. -Un Método para la Identificación Automática del Lenguaje Hablado Basado en Características Suprasegmentales Ana Lilia Reyes Herrera Doctor en Ciencias en el área de Ciencias Computacionales.- 2007.

[14] D. G. M. John G.Proakis. Digital Signal Processing. Third. New Jersey. USA: Pearson Education. 1996.

[15] F. Snell. Roy;Milinazzo. "Formant Location From LPC Analysis Data." IEEE Tansaction speech audio Process.. vol. 1. 1993.

[16] M. W. Bhatti. Y. Wang. and L. Guan. "A Neural Network Approach for Human Emotion Recognition in Speech." ISCAS. pp. 0–3. 2006.

[17] S. M. Al-qaraawi and S. S. Mahmood. "Wavelet Transform Based Features Vector Extraction in Isolated Words Speech Recognition System." Int. Symp. Commun. Syst. Networks Digit. Sign. pp. 847–850. 2014.

[18] A. L. Reyes-herrera. L. Villaseñor-pineda. M. Montes-y-gómez. and L. E. Erro. -Automatic Language Identification using Wavelets.- INTERSPEECH. 2006.

[19] S. R. Safavian and D. Landgrebe. "A Survey of Decision Tree Classifier Methodology." IEEE Trans. Syst. Man Cybern.. vol. 21. no. 3. pp. 660–674. 1991.

[20] T Kaddar. J Al- Daher. "Using Data Mining Tools For Human Resource Management" Damascus University Journal for basic Sciences. 2013.

[21] أكاديمية حسوب. (٢٠١٩). تلخيص البيانات واختصارها عبر تحليل المكونات الرئيسية (PCA) في لغة R. [online] Available at: https://academy.hsoub.com/programming/r-language/ [Accessed 12 Jun. 2019].

الباب الثالث

تحليل الآراء العربية إلكترونياً

د.أمجد يوسف أبو جبارة

تحليل الآراء العربية إلكترونياً

د.أمجد يوسف أبو جبارة

الملخص

يتناول هذا الباب استعراضاً استقصائياً لموضوع التحليل الآلي للآراء وتطبيقاته في اللغة العربية. يقدم الباب تعريفاً تفصيلياً بالموضوع يتضمن شرحاً للمهام المطلوبة لتمكين الحاسب الآلي من فهم الآراء، واستعراضاً للمقاربات البحثية المختلفة لتنفيذ كل من هذه المهام مع مقارنتها والمفاضلة بينها كلما دعت الحاجة. كما يتضمن الباب عرضاً لأهم الخوارزميات التي اقترحها الباحثون للتنقيب عن الآراء وتصنيفها مع تسليط الضوء على ما استهدف اللغة العربية منها. يتطرق الباب كذلك إلى بعض التطبيقات الرئيسية لتحليل الآراء كتصنيف مراجعات المنتجات في المتاجر الإلكترونية، والتنقيب عن الآراء في الشبكات الاجتماعية. ويختتم الباب باستعراضٍ موجز لبعض الموارد المفيدة في المجال من مجموعات نصية، ومعاجم آراء، ومكتبات برمجية.

تحليل المشاعر والآراء

تحليل المشاعر Sentiment Analysis (ويعرف أيضاً بـ«التنقيب عن الآراء» Opinion Mining) هو أحد مجالات لسانيات الحاسب الآلي Computational Linguistics المتفرعة عن مجال الذكاء الاصطناعيArtificial Intelligence ، وهو واحد من أنشط فروع هذه العلوم بحثا نظراً لأهمية تطبيقاته ووفرة المحتوى النصي اللازم لإجراء البحوث فيه، لاسيما ما تقدمه شبكات التواصل الاجتماعي اليوم من كميات مهولة من النصوص المحملة بآراء أصحابها تجاه كل أنواع القضايا التي يمكن تخيلها.

تقوم خوارزميات تحليل المشاعر بتحليل النص اللغوي بهدف الكشف عن المشاعر التي يعبر عنها الكلام تجاه موضوع النص، وبينما تركز أغلبية الخوارزميات على تصنيف المشاعر إلى إيجابية أو سلبية أو محايدة، فإن بعض الخوارزميات تذهب إلى تصانيف أشمل يتضمن حالات شعورية أكثر تفصيلاً كالسعادة والحماسة والغضب والاشمئزاز، إلخ.

ولعل أهم العوامل التي ساعدت في نشأة وتطور هذا العلم هو تطبيقاته المهمة في مجالات التسويق، وخدمة العملاء، وتطوير المنتجات، وقياس الرأي العام، والعلوم السياسية، والدراسات الاجتماعية، وغيرها الكثير. حتى أصبح تحليل الآراء خدمة مدفوعة تقدمها شركات متخصصة وتستفيد منها جهات عديدة (من شركات ومنظمات وحكومات) معنية برصد وقياس آراء زبائنها أو مستخدمي منتجاتها أو المستفيدين من خدماتها.

نبذة تاريخية

تعود أصول مجال تحليل الآراء والمشاعر إلى علم الفلسفة، وتستند الكثير من الدراسات الأولى في الموضوع إلى أفكار فريدريك نيتشه ونظرياته حول تعدد الآراء Perspectivism التي تتلخص في أن الحقيقة ممكن أن تكون ذات أوجه متعددة، وأن كثير من القضايا التي يتجادل حولها الناس ليس لها حقيقة مطلقة بالضرورة [١] [٢] [٣].

الفيلسوف الأمريكي ريتشارد سكاشت درس أفكار نيتشه، وأعاد صياغتها بحيث فرق بين نوعين من الأفكار: الأفكار المرتبطة بحقائقObjective ، والأفكار التي تعبر عن رأي Subjective [٤]. وتعتبر هذه الدراسات هي الأرضية التي ارتكزت عليها الكثير من الدراسات الحديثة في مجال تحليل الآراء.

ولما كان الكلام المكتوب والمنطوق هو الوسيلة الرئيسية للتعبير عن الأفكار ومشاركتها مع الآخرين، فقد انصب كثير من اهتمام الباحثين في هذا المجال على دراسة العلاقة بين طبيعة الكلام المستخدم في الحديث والآراء التي يحملها المتحدث [٥] [٦] [٧]، حتى ظهر مجال في علم اللغويات متخصص بدراسة اللغويات النفسية Psycolinguistics [٨]. فعلى سبيل المثال درست الباحثة آن بانفيلد Ann Banfield الجمل التي تعبر عن الحالة النفسية للمتحدث من حيث كونه يسرد حقائق موضوعية أو يعبر عن آراء، وعلاقة ذلك باختيار الألفاظ والتعبيرات وتركيب الجمل [٩]، كما ظهر مجالٌ أكثر تخصصاً يتعلق باللغويات الاجتماعية Sociolingistics [١٠]، ويهتم بدراسة الطرق المختلفة التي يستخدمها الناس للتعبير عن أفكارهم في أوضاع التفاعل الاجتماعي المختلفة كحال الاتفاق أو الإعجاب أو المعارضة إلخ.

وقد مثلت كل هذه الدراسات المختلفة أساساً بنى عليه الباحثون المهتمون بمجال معالجة اللغات مقارباتهم approaches المختلفة لبناء أنظمة حاسوبية قادرة على تحليل الآراء التي يتم التعبير عنها بطريق الكلام. ومن الرواد في هذا المجال الباحثة جينيس ويب Janyce Wiebe التي استفادت من دراسة بانفيلد سابقة الذكر لتطوير خوارزمية قادرة على اكتشاف أنماط الكلام التي تظهر بشكل متكرر مع الحالات النفسية وفي الحالات الاجتماعية المختلفة [١١]. ومن أمثلة الدراسات الريادية المهمة في هذا المجال كذلك ما قام به الباحث ستيفن جرين من تطوير خوارزميات قادرة على كشف أنماط الكلام التي تعبر عن ميول وتحيزات ضمنية لا يتم التعبير عنها بشكل صريح في الكلام، وقد تضمن بحثه إجراء دراسات لغوية اجتماعية واقعية متعددة لتدعيم استنتاجاته واختبار دقة خوارزميته [١٢].

ومن أوائل التطبيقات العملية الحديثة التي انصب عليها تركيز باحثي لغويات الحاسب الآلي فيما يتعلق بتحليل الآراء: أنظمة إجابة الأسئلة Question Answering Systems. وكانت بؤرة التركيز فيها هي تطوير هذه الأنظمة بحيث تصبح- إلى جانب قدرتها على إجابة الأسئلة المرتبطة بحقائق- قادرة كذلك على إجابة أسئلة الرأي التي تحتمل أكثر من إجابة.

وكان من أهم الجهود الريادية في هذا المجال ما قامت به الباحثة جينيس ويب عام ٢٠٠٢ عندما نظمت ورشة عمل استمرت شهرين جمعت فيها عدداً من الباحثين لدراسة كيفية استخدام الناس للغة للتعبير عن الآراء. وخرجت هذه الورشة بمجموعةٍ من التعريفات المحددة التي تميز الكلام الحمال للرأي عن الحقائق، ومعايير تصنيف الكلام الحمال للرأي إلى كلام إيجابي أو سلبي أو محايد. كما قام المشاركون في هذه الورشة بتطبيق هذه التعريفات والمعايير على مدونة نصية Text Corpus مأخوذة من مقالات إخبارية لتشكل هذه المجموعة ما يعرف الآن بـ MPQA والتي أصبحت أحد أهم المجموعات النصية التي يستخدمها باحثو لغويات الحاسب الآلي لتدريب واختبار خوارزميات تحليل الآراء [١٣].

ومع ظهور وانتشار مواقع التجارة الإلكترونية وإقبال الناس المتزايد على شراء احتياجاتهم عبر الإنترنت، ومع ما تقدمه هذه المواقع في الغالب للمشترين من إمكانية

التعليق على المنتجات التي قاموا بشرائها وتبيان ما أعجبهم وما لم يعجبهم فيها، انصب اهتمام باحثي تحليل المشاعر والآراء على دراسة هذه التعليقات واقتراح خوارزميات تسهل على الباعة والمصنعين معرفة مقدار إعجاب الناس بمنتجاتهم مع تلخيص الجوانب التي لاقت استحسان المشترين والجوانب التي طالها نقدهم [١٤] [١٥] [١٦][١٧].

ثم مع ظهور وانتشار مواقع الإعلام الاجتماعي والشبكات الاجتماعية، توفرت ميادين واسعة لمستخدمي الإنترنت للتعبير عن آرائهم تجاه كل القضايا، بل والخوض في جدالات حول مواضيع الاختلاف سواءً كانت هذه المواضيع تقنية أو فكرية أو سياسية [١٨]. استقطبت هذه الوفرة المهولة في النصوص الحمّالة للآراء جهوداً بحثيةً كثيرةً انصب جُلُّ اهتمامها على محاولة فهم اللغة التي يستخدمها الناس للتعبير عن آرائهم عبر وسائل التواصل الاجتماعي، والمفردات والتعبيرات التي يستعملها الناس في كلامهم حال الاتفاق أو الاختلاف، وكيف يمكن استخدام تقنيات معالجة اللغات لتحليل النصوص الحمالة للآراء بهدف تصنيفها آلياً وكشف علاقات الاتفاق والاختلاف بين أصحابها [١٩][٢٠][٢١].

ومن تطبيقات تحليل الآراء الأخرى التي لاقت اهتماماً متزايداً في السنوات الأخيرة دراسة طرائق التعبير عن الآراء في السياق الأكاديمي، وتحديداً عندما يشير الباحثون إلى أعمال باحثين آخرين ويتعرضون لها بالنقد. يحاول الباحثون في هذا المجال إحداث نقلة في معايير تقييم المساهمات العلمية للباحثين بحيث لا يتم الاكتفاء بتعداد الإشارات المرجعية التي يتلقاها العمل البحثي، بل يتم النظر أيضاً إلى طبيعة الرأي المصاحب للإشارة وهل هو رأي مؤيد أم معارض لما جاء به البحث المشار إليه [٢٢][٢٣][٢٤] [٢٥][٢٦].

تحليل الآراء العربية

جهود البحث في تحليل الآراء العربية جاءت متأخرة نوعاً ما، بعد أن وفر انتشار وسائل التواصل الاجتماعي وتعاظم أثرها عربياً وعالمياً حافزاً كبيراً لدى كثير من الباحثين من عرب وغيرهم لمباشرة البحث في هذا المجال. ركزت الجهود الأولى على

مواءمة المقاربات المستخدمة لتحليل الآراء في اللغة الإنجليزية واللغات الأخرى للغة العربية، وتضمن هذا بناء موارد لغوية تخدم تحليل الآراء العربية كمعاجم آراء ومدونات لغوية Corpora مصنفة يدوياً ومكتبات برمجية لتحليل الآراء [٢٧][٢٨][٢٩][٣٠]. انتقلت الجهود البحثية في هذا المجال بعد ذلك إلى التعامل مع التحديات الخاصة باللغة العربية كتعدد اللهجات العربية [٣١][٣٢][٣٣]، ودراسة أثر المعالجة المسبقة للنص العربي (كالتحليل الصرفي والتجذير والتجذيع) على دقة تحليل الآراء.

نُشرت العديد من الأبحاث الاستقصائية في السنوات الأخيرة حول تحليل الآراء في اللغة العربية ولخصت الجهود البحثية في المجال على اختلاف محاور تركيزها وتطبيقاتها والطرق التي استخدمتها والتحديات التي عالجتها، وندعو القارئ المهتم إلى الرجوع إلى هذه الدراسات كقراءة مكملة لما يحتويه هذا الباب [٣٤][٣٥][٣٦][٣٧][٣٨][٣٩][٤٠].

المهام الرئيسية في تحليل الآراء

نستعرض في هذا القسم العمليات والمهام المختلفة التي تصدى لها الباحثون في مجال تحليل الآراء، ونكتفي هنا بتعريف هذه المهام والإشارة إلى أهم الأبحاث التي تصدت لكل منها، الشرح الأكثر تفصيلاً لطرق إجراء هذه المهام سنتطرق إليه في القسم التالي.

- **تمييز الكلام الحمال للآراء**

وتعتبر هذه المهمة (ويشار إليها في الأبحاث عادة بـ«تحليل موضوعية الكلام» (Subjectivity Analysis) بمثابة المهمة الأساسية الأولى في معظم عمليات تحليل الآراء، وتستند الأبحاث الأولى فيها إلى الدراسات اللغوية النفسية والفلسفية والاجتماعية كما أشرنا آنفاً.

الهدف من هذه المهمة هو التمييز بين الكلام الذي ينقل حقائق والكلام الذي يعبر عن رأي، فمثلاً قول أحدهم: «كشفت شركة سامسونج النقاب عن هاتفها الجديد يوم الخميس الماضي» إنما ينقل خبراً يتعلق بهاتف سامسونج دون التعبير عن أي رأي أو أي مشاعر مرتبطة بهذا الحدث أو موجهة تجاه الهاتف الجديد. قارن هذا بـ: «الهاتف الجديد الذي أعلنت عنه سامسونج رائع، وفيه الكثير من الخصائص المميزة»، فالكلام في هذه

الحالة يعبر عن رأي صاحبه المتحمس للهاتف الجديد وما به من خصائص يراها مميزة.

وغالباً ما يجرى هذا النوع من التحليل على مستوى الجمل، حيث يتم تصنيف كل جملة في النص إلى جملة موضوعية Objectiveأو جملة معبرة عن رأي Subjective اعتماداً على ما تحويه الجملة من ألفاظ [٤١] [٤٢]. فالجمل الحمالة للرأي تتميز باحتوائها على صفات (إيجابية أو سلبية) مثل «رائع» و«المميزة» كما في المثال السابق، في حين أن الجمل الموضوعية تحتوي غالباً على أرقام أو تواريخ أو غيرها من التعبيرات التي يكثر اقترانها بنقل الحقائق أو توثيق الأحداث.

وإذا لزم تصنيف موضوعية نص كامل فإن ذلك يتم بطريقة إحصائية في الغالب من خلال رصد موضوعية الجمل المكونة للنص، فكلما زادت نسبة الجمل الحمالة للرأي في النص، اعتبر النص في مجمله أكثر ميلاً نحو كونه نصاً معبراً عن رأي والعكس صحيح.

- **تحديد قطبية الكلام**

بعد تحديد الكلام الحمال للرأي تأتي المهمة التالية وهي التعرف على نوعية المشاعر التي يعبر عنها النص. الغالبية الراجحة من الدراسات ركزت على تصنيف المشاعر إلى مشاعر سلبية ومشاعر إيجابية مع إمكانية التمييز بين درجات مختلفة من قوة أو ضعف الإيجابية أو السلبية. ويطلق على الخاصية التي تصف الكلام من حيث كونه سلبياً أو إيجابياً في الأوساط البحثية بـ«قطبية الكلام» Text Polarity، وتعرف أيضاً بـ«الانحياز المعنوي» Semantic Orientation.

تطرقت أبحاث تحليل قطبية الكلام إلى دراسة القطبية على مستويات مختلفة ابتداءً من قطبية الكلمات وصولاً إلى قطبية النصوص الكاملة.

- **تمييز قطبية الكلمات:**

وتهدف هذه العملية إلى تصنيف الكلمات الواردة في النص إلى كلمات إيجابية (مثل: جميل، حسن، رائع، كريم، إلخ) أو كلمات سلبية (مثل: سيء، رديء، هزيل، بخيل، إلخ) أو كلمات محايدة (مثل: ذَهَبَ، مَع، كِتاب، شارِع، إلخ). للوهلة الأولى قد تبدو هذه العملية سهلة وأن الكلمات السلبية والإيجابية يمكن حصرها في معجم حصراً يدوياً (وهو ما قام به العديد من الباحثين في مجال اللغويات النفسية والاجتماعية بالفعل

[٤١] [٤٣] [٤٤] [٤٥])، ولكن هذه العملية في الحقيقة تحيط بها تحديات متعددة تجعل المعاجم اليدوية غير قادرة على تلبية احتياجات معظم تطبيقات تحليل الآراء:

- فالمعاجم اليدوية المتاحة مهما كَبُرَت أحجامها تظلُّ عاجزةً عن حصر كل الكلمات التي تحمل دلائل قطبية، خاصة أن كثيرا من تطبيقات تحليل الآراء تجري على نصوص منشورة على الإنترنت حيث تظهر مفردات جديدة باستمرار للتعبير عن معاني سلبية أو إيجابية (مثل وصف الأفكار بأنها «داعشية» - وهو لفظ مستحدث لا تحويه معاجم القطبية)، ويغلب استعمال الكلام العامي، ويكثر استعمال الاختصارات (مثل استعمال gr8 كاختصار ل great)، واستعمال الوجوه التعبيرية، وغيرها.

- كما أن معاجم القطبية متوفرة لعدد محدود من اللغات فقط، في حين أن عدد كبير من اللغات لا توجد لها معاجم قطبية على الإطلاق أو أن ما هو متوفر منها يعاني من محدودية المحتوى وغياب الاهتمام بتحديثه.

- كذلك توجد كلمات كثيرة تحتمل معانٍ متعددة، ويختلف معناها بحسب السياق، وبناءً على المعنى المقصود قد تتنقَّل قطبيتها بين إيجابية وسلبية ومحايدة، فمثلاً كلمة «أسد» في معناها الغالب هي اسم حيوان مفترس، ولكن في سياقات معينة تكون لها دلالة إيجابية كقولهم «أنت أسد» في استعمالٍ مجازي يرُاد منه التعبير عن صفات الشجاعة والقوة. تتجنب المعاجم القطبية إدراج هذه الكلمات لأن الغالب عليها هو المعنى المحايد، في حين أن كثير من تطبيقات تحليل الآراء تحتاج إلى أن تكون قادرة على التعرف على المقصد القطبي لهذه الكلمات.

- أخيراً، تختلف الكلمات القطبية في مقدار قطبيتها، فكلمة «ممتاز» -مثلاً- تتعبر أقوى في دلالتها الإيجابية من كلمة مثل «جيد». مثل هذا التقدير لدرجة الإيجابية أو السلبية غير متاح في الغالبية العظمى من المعاجم القطبية، وما هو موجود منها يكتفي بتصنيف قطبية الكلمات إلى قوية وضعيفة فقط.

بسبب هذه التحديات ومحدودية المعاجم اليدوية انصب اهتمام كثير من الباحثين الأوائل في مجال تحليل الآراء على البناء الآلي للمعاجم أو الإثراء الآلي للمعاجم اليدوية الموجودة، واستخدم الباحثون طرقاً متعددة لتحقيق هذا الهدف نستعرض بعضاً منها في الفقرات التالية.

استندت كثير من هذ الطرق إلى فرضية أن الكلمات التي تحمل دلالات قطبية متشابهة تظهر غالباً في مواضع متقاربة، فمثلاً إذا كان هناك نص يبدي رأياً تجاه منتج جديد، وإذا كنا نعرف قطبية بعض الكلمات الواردة في هذا النص، فيمكن افتراض أن باقي الصفات الواردة في النص من الممكن أن تحمل قطبية مماثلة، وإذا أجرينا هذا الرصد للظهور المتزامن لكلمات معروفة القطبية مع بقية الكلمات على كمية ضخمة جداً من النصوص يصبح من الممكن رصد علاقات اقتران إحصائية تقود إلى تخمين قطبية الكلمات غير معروفة القطبية. فمثلاً الكلمات التي تتكرر على مقربة من كلمات معروفة الإيجابية يمكن افتراض أنها إيجابية، والأمر كذلك مع الكلمات التي تتكرر مع كلمات سلبية، أما الكلمات التي ترد بنفس مقدار التكرار مع كلمات إيجابية وكلمات سلبية فيمكن افتراض أنها كلمات متعادلة القطبية [٤٦] .

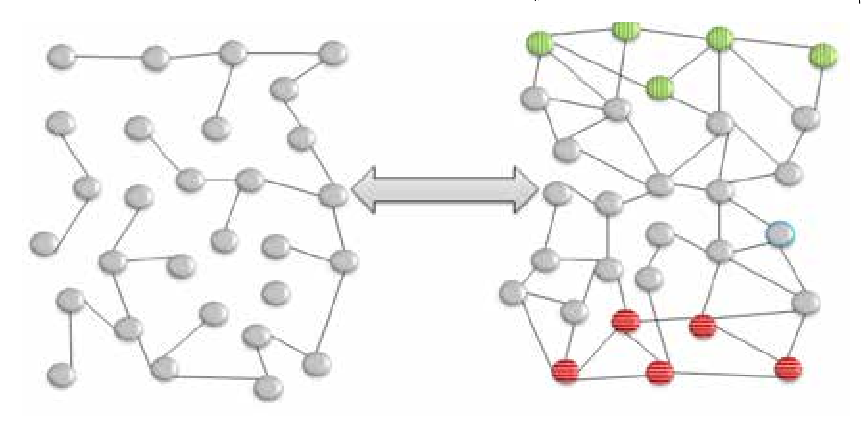
حاولت مقاربات أخرى النظر إلى الطريقة التي ترتبط فيها الصفات التي تتجاور في النصوص وتفصلها حروف عطف أو حروف استدراك أو ما شابه، ومحاولات استنتاج القطبية للكلمات مجهولة القطبية بمساعدة الكلمات ذات القطبية المعروفة. فمثلاً إذا احتوى نص على شيء من قبيل: «جميل ورائع» وكانت قطبية «جميل» معروفة مسبقاً، فإن حرف العطف «و» يوفر قرينة قوية بأن كلمة «رائع» تحمل نفس القطبية. أما إذا احتوى نص على تعبير مثل: «جميل لكنه مزعج»، وكانت قطبية «جميل» معروفة، فإن حرف الاستدراك «لكن» يمنح قرينة قوية بأن كلمة «مزعج» لها قطبية معاكسة [٤٧].

عمدت طرق أخرى إلى الاستفادة من شبكات الكلمات Word Networks، وهي شبكات تكون كل نقطة node فيها عبارة عن كلمة، وترتبط الكلمات ببعضها بروابط edges تمثل علاقات ترادف أو تضاد أو غيرها من العلاقات المعنوية Semantic Relationships. الطرق التي تعتمد على هذه الشبكات تستخدم خوارزميات التعلم الآلي شبه الموجه Semi-supervised learning للتعرف على قطبية الكلمات المختلفة

في الشبكة انطلاقاً من عدد قليل -نسبيا- من الكلمات معروفة القطبية يتم اختيارها يدوياً [٤٨] .

من هذه الخوارزميات ما يعتمد على التنقل العشوائي في الشبكة Random Walks، ولتحديد قطبية كلمة ما باستخدام هذه الطريقة فإن عملية التنقل العشوائي تنطلق من تلك الكلمة وتستمر في التنقل العشوائي عبر الشبكة حتى تصل إلى كلمة معروفة القطبية، ويتم تكرار هذه العملية مرات كثيرة لكل كلمة، وفي النهاية يتم تعيين قطبية للكلمة بحسب القطبية التي غلبت على الكلمات ذات القطبية المعروفة التي توقفت عندها عملية التنقل في كل محاولة. أما في حال تعذر إيجاد أغلبية واضحة لإحدى القطبيتين فيتم اعتبار أن الكلمة ذات قطبية متعادلة [٤٩].

حاولت مقاربات أخرى إثراء المعاجم القطبية للغات التي تعاني من فقر المعاجم وفقر الموارد النصية التي تتيح بناء معاجم آلية لها (كقلة المحتوى المكتوب بتلك اللغة عبر الإنترنت مثلاً) من خلال الاستفادة من معاجم لغات أخرى تتميز بثراء معاجمها، ومن هذه الطرق مثلاً ما يعمد إلى بناء شبكات كلماتٍ متعددة اللغات Multi-lingual Word Networks من خلال استخدام القواميس وربط الكلمات بترجماتها من اللغات المختلفة. يتبع ذلك استخدام خوارزميات كالتي عرضناها في الفقرة الماضية لاستنتاج قطبية الكلمات غير معروفة القطبية في اللغات المختلفة انطلاقاً من بعض كلمات معروفة يتم اختيارها يدوياً، كما هو مبين في شكل ١ [٥٠].



شكل ١ يوضح شبكتين WordNet للغتين مختلفتين، الأولى -يمين- تحتوي على كلمات معروفة القطبية، والثانية -يسار- تخلو من هذه المعلومات ولكنها مرتبطة بالشبكة الأخرى من خلال ترجمة الكلمات

- **التعرف على قطبية الجمل والفقرات**

المهمة التي تحدثنا عنها في القسم السابق تهتم بدراسة قطبية الكلمة بشكل مجرد معزول عن سياقها الذي وردت فيه. في هذا القسم سنتحدث عن مهمة أكثر تعقيداً وهي التي يؤخذ فيها السياق بعين الاعتبار، وهي خطوة مهمة لأن السياق له دور كبير في تحديد قطبية الكلمة، ونستعرض فيما يلي بعض الحالات التي يؤثر فيها السياق على الكلمات مع ذكر أمثلة على كل منها.

- بعض الكلمات تحتمل أكثر من معنى. فقد تستخدم الكلمة في سياق فتحمل معنىً إيجابياً وقد تستخدم في سياقٍ آخر فتحمل معنىً سلبياً أو تكون محايدةً، ومثال ذلك كلمة «أسد» كما أوردنا سابقاً. مثال آخر كلمة «عين»، فقد تأتي بمعنىً محايد كما في: «اشتريت قطرة عينٍ لعلاج الاحمرار»، أو بمعنىً إيجابي عندما تستخدم استخداماً مجازياً كما في «ابني هو عيني ولا غنى لي عنه»، أو بمعنىً سلبي كما في «كان عيناً للأعداء» أي «جاسوساً».

- إذا وردت الكلمة القطبية في سياق نفي فإن قطبيتها تنعكس. فمثلاً في جملة: «لا أحب الباذنجان» الأصل في كلمة «أحب» أنها موجبة القطبية، ولكن ورود حرف «لا» في بداية الجملة، ووقوع كلمة «أحب» في نطاق نفيها، قلب قطبيتها من موجبة إلى سالبة.

- قد ترد الكلمة القطبية في سياق نفي ولكن لا يؤدي النفي إلى عكس قطبيتها بالضرورة، ولكن يؤدي إلى التقليل من قوة قطبيتها Sentiment Intensity، فمثلاً في جملة «لا أحب الباذنجان كثيراً» برغم أن كلمة «أحب» وردت في سياق النفي، إلا أن تذييل الجملة بـ«كثيراً» قد جعل المنفي هو كثرة المحبة وليس أصلها.

- قد ترد الكلمات القطبية في سياق السخرية ويكون مقصد قائلها معاكساً لقطبيتها الظاهرة. فمثلاً قد يقول أحد للآخر «يا ذكي» في سياق من السخرية يكون مقصده فيه أن الموجه إليه الكلام قليل الذكاء وهو ما يعاكس ظاهر المعنى. ويعتبر التعامل مع حالات السخرية في الكلام من أصعب مشكلات تحليل الآراء، وذلك لأن تمييز الكلام الجاد من الكلام الساخر يحتاج في أغلب

الأحيان إلى معرفة الثقافة اللغوية السائدة بين المتحدثين، وهو ما يتجاوز كثيراً نطاق النص الذي يجري تحليله.

للتعامل مع هذه التحديات التي تستوجب إدراك السياق حاولت بعض طرق تحليل الآراء استخدام بعض الخوارزميات المبنية على قواعد مصاغة يدوياً، مثلاً في حال ورود كلمة نفي في الجملة يتم عكس قطبية كل الكلمات القطبية الواردة في نفس الجملة وعلى بعد مسافة محددة من أداة النفي وهكذا، ولكن هذه الطرق تعاني من عدم مقدرتها على اكتشاف كل أنواع السياق المؤثرة في قطبية الكلام. ولذلك فإن الكثير من طرق تحليل الآراء قد اعتمدت على تقنيات تعلم الآلة Machine Learning سواءً الطرق التقليدية منها أو طرق التعلم العميق Deep Learning. في حالة طرق تعلم الآلة التقليدية ينصب جهد الباحثين على تعريف إشارات وخصائص Features ممكن إيجادها في النص ويمكن أن يكون لها أثر في قطبية الكلام، ومن أمثلة هذه الخصائص ما يلي:

- الكلمات المجاورة (الكلمة السابقة والتالية مثلاً) للكلمات القطبية في الجملة.
- وجود أداة نفي في الجملة، والمسافة -مقاسة بالكلمات- بين أداة النفي والكلمات القطبية في الجملة.
- وجود كلمات تقوية Intensifiers أو تضعيف Downtoners مقترنة بالكلمة القطبية مثل: «بشدة»، «بقوة»، «كثيراً»، «جداً»، «قليلاً»، إلخ.
- العلاقات الإعرابية بين الكلمات في الجملة، لاسيما بين الكلمة القطبية وغيرها من الكلمات كأدوات النفي أو كلمات التقوية والتضعيف وغيرها.
- احتواء الجملة على وجوه تعبيرية Emoticons، أو علامات ترقيم (مثل علامة تعجب أو علامة استفهام)، أو رموز تزينية، أو وسوم تصنيفية Hashtags، أو التطويل لبعض الحروف في بعض الكلمات كما في «عجيـــــــــــــــــــــب»، أو تكرار الحروف كما في «رااااااائع»، إلخ.

هذه الخصائص يتم تعريفها لكل جملة أو فقرة في النص، وعند توفر كمية كافية من الجمل أو الفقرات معروفة القطبية، يتم تدريب خوارزميات تعلم الآلة على هذه الأمثلة، حتى تصبح قادرة على تخمين قطبية أي جمل أو فقرات أخرى.

مؤخراً -ومع الوفرة الكبيرة للبينات المحملة بالآراء المنشورة على الإنترنت- شهدت تقنيات تحليل الآراء صعود الطرق المعتمدة على التعلم العميق، وفيها ينصرف تركيز الباحثين عن تعريف خصائص صريحة لاكتشاف القطبية في ضوء السياق إلى التركيز على بنية النموذج العميق Model Architecture الذي يراد تدريبه. وسوف نتحدث بقدر أكبر من التفصيل عن هذه الطرق لاحقاً في هذا الباب.

- **التعرف على مصدر الرأي**

كثير من تطبيقات تحليل الآراء تهتم بتمييز الآراء المنقولة عن آخرين. فمثلاً عند قول أحدهم: «صديقي يكره منتجات شركة آبل، ولكني أحبها»، نجد أن التعبير السلبي «يكره» ليس مقترناً بالمتحدث صاحب النص، وإنما هو ينقل مشاعر مصدرها مختلف. ولذلك فإن طرق تحليل الرأي في مثل هذه التطبيقات تحتاج إلى ربط كل تعبير قطبي في النص بمصدره والتمييز بين كونه مقترناً بالكاتب أم بمصدر آخر.

- **التعرف على المستهدف بالرأي**

كثير من تطبيقات تحليل الآراء تهتم أيضاً برصد رأي صاحب النص تجاه منتج محدد أو خدمة محددة، ولذلك يلزم معرفة المستهدف بكل تعبير قطبي في النص. فمثلاً إذا قال أحدهم: «أنا أحب هواتف آبل، ولكني أكره أجهزتها اللوحية»، تحتاج أكثر تطبيقات إلى القدرة على تمييز أن مشاعر المحبة موجهة للهواتف، بينما مشاعر الكره موجهة نحو اللوحيات وليس العكس.

كذلك فإن كثير من تطبيقات تحليل الآراء تتطرق إلى رصد رأي الناس في خدمات أو منتجات متعددة الجوانب، ومن الممكن أن يختلف تقييم الناس لكل من هذا الجوانب، فمثلاً عند قيام المستخدمين بتقديم تقييم نصي لأحد المطاعم فإن هذا التقييم قد يتطرق إلى جودة الطعام، ترتيب ونظافة مكان الجلوس، لباقة النادل، الأسعار، إلخ. فمثلاً في تعليق مثل: «الطعام لذيذ جداً، وتعامل طاقم المطعم راق، ولكن الضوضاء في المكان شديدة والإضاءة ضعيفة» نجد خليطا من آراء إيجابية وسلبية. وتحتاج كثير من تطبيقات.

ونظراً لأهمية هذا الربط بين الرأي والجانب المستهدف بالرأي لأكثر تطبيقات تحليل الآراء فقد ظهر مجال خاص يعرف بـ«تحليل الآراء متعدد الجوانب.» Aspect-based Sentiment Analysis.

وتعتمد الطرق التقليدية المهتمة بمعرفة مصدر ووجهة الرأي على تحليل العلاقات الإعرابية في الجملة بين الكلمات القطبية والكلمات الأخرى لاسيما الجمل الاسمية Noun Phrases والكيانات المسماة Named Entities. أما طرق التعلم العميق فتحاول كشف العلاقات الإعرابية بشكل ضمني من خلال بنية النموذج Model Architecture الذي يتم تدريبه دون أن يتم إجراء عملية الإعراب نفسها بالضرورة.

- **مهام متقدمة لتحليل المشاعر**

المهام التي تناولناها في الفقرات السابقة تعتبر مهام أساسية ولازمة للغالبية العظمى من تطبيقات تحليل الآراء. نتناول هنا على عجالة بعض المهام المتقدمة التي قد تحتاجها بعض تطبيقات تحليل الآراء.

- **تلخيص الآراء**

كما ذكرنا سابقاً فإن العديد من تطبيقات تحليل الآراء تتعامل مع حالات تتعدد فيها الجوانب التي يستهدفها الناس بآرائهم، مثل تعليق الناس على أحد المنتجات كهاتف مثلاً فيستحسنون جودة الكاميرا مثلاً ولكنهم يتضجرون من قصر عمر البطارية أو يعجبهم الشكل الأنيق للهاتف ولكن يضايقهم تأخر استجابة شاشة اللمس وهكذا.

في هذه التطبيقات لا يكفي وسم تعليق المستخدم بأنه إيجابي أو سلبي بمجمله بل يجب تفصيل الجوانب الإيجابية والجوانب السلبية من وجهة نظر كل مستخدم.

تهدف مهمة تلخيص الآراء إلى تصنيف الآراء المختلفة للمستخدمين من حيث الجوانب التي استهدفتها آراؤهم، بحيث يتم وضع الآراء الخاصة بكل جانب في مجموعة واحدة ثم يتم تصنيفها إلى إيجابية وسلبية. ثم يتم تطبيق آليات تلخيص النصوص Text Summarization على مجموعة النصوص الخاصة بكل منها، ويكون المخرج النهائي لهذه العملية هو ملخص مفصل يعرض كل جانب على حدة وأهم الآراء الإيجابية والسلبية التي استهدفت كل جانب.

- **تتبع تطور الآراء**

يحاول الباحثون المهتمون بهذا النوع من تحليل الآراء دراسة الطبيعة الديناميكية للآراء وتتبع تطورها وتغيرها مع الوقت. ففي حالة تحليل آراء المستفيدين من خدمة ما -مثلاً-، قد يكون من المفيد تتبع التغير الذي يطرأ على آرائهم بعد إجراء أي تغييرات في الخدمة، وملاحظة كيف تميل الآراء نحو الإيجابية أو السلبية كردة فعل من طرف المستفيدين.

كذلك في مجموعات النقاش عبر الشبكات الاجتماعية، تهتم العديد من الدراسات الاجتماعية برصد كيف يؤثر سير النقاش على آراء المشاركين فيه وإذا ما كان أحدهم سيغير رأيه مع مرور الوقت، وتأثير سير النقاش كذلك على الرأي المبدئي الذي يتبناه من ينخرط في النقاش متأخراً.

- **رصد انقسام مجموعات النقاش حول موضوع النقاش**

من مجالات الدراسة التي يعنى بها الباحثون في مجال تحليل الآراء دراسة انقسام المنخرطين في نقاشات جدلية حول موضوع النقاش، ودراسة اللغة التي يستخدمونها في التعبير عن انقسامهم. ويتم تطبيق هذه الدراسات غالباً على الحوارات التي تحوي العديد من منشورات الأخذ والرد بين المشاركين في النقاش كما في منتديات الحوار وغيرها من وسائل التواصل الاجتماعي؛ فيحاول الباحثون تحليل المنشورات التي يكتبها كل مشارك وتحديد ما إذا كانت تعبر عن اتفاق أو اختلاف مع رأي المنشور السابق الذي جاءت رداً عليه، فمثلاً إذا بدأ المشارك تعليقه على منشور سابق بقوله: «هذا رأي خاطئ» أو «أنا أختلف مع هذا الرأي» أو ما شابه ذلك تحاول هذه الخوارزميات أن تستنج أن صاحب الرد وصاحب التعليق الأصلي على طرفي نقيض فيما يتعلق بموضوع النقاش. وتذهب الدراسات إلى أبعد من ذلك فهي تحاول كذلك أن ترصد مواطن الاتفاق والاختلاف بين المتحاوريين، فقد يختلف متحاورين حول أحد جوانب النقاش ولكن قد يختلفان في جانب آخر من مثل: «أتفق معك في كذا، ولكني أخالفك الرأي في كذا». تحاول الأبحاث في هذه الحالة بناء «سجل انطباعات» Attitude Profile لكل مستخدم تسجل فيه انطباعات المستخدم السلبية أو الإيجابية تجاه المستخدمين الآخرين وتجاه الجوانب المختلفة لموضوع النقاش.

تستند كثير من هذه الدراسات إلى نظريات في العلوم الاجتماعية كنظرية التوازن البنائي Structural Balance Theory والتي ترصد ظواهر اجتماعية متكررة تفسر انقسام الناس حول الآراء المختلفة مثل «صديق صديقي صديقي» و«عدو عدوي صديقي»، وهكذا.

- **رصد التأثير على الآراء في المناظرات**

تحاول الدراسات المهتمة بهذا الجانب رصد عملية التأثير التي تجري في الحوارات التي تدور عبر منصات الحوار الإلكتروني كالشبكات الاجتماعية وما شابهها، والتعرف على الأشخاص المؤثرين الذين يوجهون سير النقاش ويؤثرون في آراء غيرهم من المشاركين وربط هذا بما لديهم من قوة اجتماعية Social Power، وسلطة على الآخرين Social Authority.

- **تصنيف أكثر تفصيلاً للمشاعر**

ثمة فرع من تحليل الآراء يتجاوز تصنيف الآراء لسلبية وإيجابية ويقترح تصنيفات أكثر تفصيلاً تتضمن مشاعر مثل الغضب، والحزن، والملل، والسعادة، والحماسة، إلخ. ولكن الدراسات في هذا الجانب ما زالت قليلة نسبياً نظراً لقلة البيانات المتاحة التي يتوفر فيها نصوص مكتوبة مقرونة بمشاعر تفصيلية.

طرق تحليل الآراء

في هذا الباب نستعرض المقاربات المختلفة التي استعملها باحثو تحليل الآراء لإجراء المهام التي عرضنا بعضاً منها في الجزء السابق من هذا الباب.

نبدأ بعرض عمليات المعالجة المسبقة Preprocessing التي يلزم القيام بها قبل البدء بعمليات تحليل الآراء مع التركيز هنا على ما تحتاجه اللغة العربية. يتبع ذلك استعراض لثلاثة مدارس في تحليل الآراء مع تقديم أمثلة لكل منها وعقد المقارنات بينها كلما قضت الحاجة.

- **المعالجة المسبقة للنصوص**

هي خطوة مهمة يجب إجراؤها قبل البدء بتحليل الآراء، خاصةً عند التعامل مع اللغة العربية، وقد بينت الدراسات أن هذا النوع من المعالجة له أثر واضح في دقة عمليات تحليل الآراء التي تتبعها [٥١]. وتعود الأهمية الخاصة لإجراء هذه المعالجة لنصوص اللغة العربية لما تتميز به من ثراء المفردات، وكثرة أشكال الصرف، وغياب التشكيل من معظم النصوص العربية المكتوبة مع ما يخلقه هذا من غموض لمعاني بعض الكلمات، وتعدد اللهجات العربية، وغيرها. وتتضمن عمليات المعالجة المطلوبة في اللغة العربية ما يلي:

- **تقطيع الكلام (إلى كلمات أو وحدات نصية) Tokenization:**

وتسمى هذه العملية أيضاً بالتحليل اللفظي Lexical Analysis ويقصد به تقطيع النص إلى وحدات Tokens تتكون كل وحدة منها من أحرف أو أرقام أو رموز متصلة كالكلمات أو الأعداد أو علامات الترقيم، مع تحديد موضع بداية ونهاية كل وحدة.

- **تسوية الكلام Orthographic Normalization**

وتهدف إلى تنقية النص من الشوائب الكتابية كالرموز الزائدة وعلامات الترقيم غير الهامة لعملية المعالجة مثلاً، والتأكد من توحيد الأنماط المختلفة لكتابة الشيء الواحد (مثل إثبات أو ترك رسم الهمزة في الألف المهموزة)، والتخلص من التطويل، والتخلص من الحروف المكررة كما في «راااائع»، وإزالة التشكيل إذا كان غير لازماً في عمليات المعالجة التالية أو غير متوفر بشكل شامل لكل النص المكتوب.

وقد بينت بعض البحوث المتعلقة بمعالجة اللغة العربية أن إجراء عمليات التسوية على النصوص العربية له تأثير ملحوظ على جودة وكفاءة عمليات المعالجة اللاحقة للنص [٥٢].

- **التحليل الصرفي Morphological Analysis**

وتهدف عملية التحليل الصرفي للكلمات إلى دراسة بنية الكلمة بغرض التعرف على القسم الصرفي للكلمة، كتحديد هل هي جمع أم مفرد، صيغة تذكير أم تأنيث، صيغة ماضٍ أم مضارع أم أمر للأفعال ... إلخ، كما تهدف إلى تحديد جذر الكلمة وتحديد الزوائد التي أدخلت على الجذر لصرفه.

وهذه العملية مهمة جداً لتحليل الآراء ففي حال الاعتماد على المعاجم القطبية لإجراء عملية التحليل فإن المعاجم المتاحة لا تحوي كل أشكال الصرف للكلمة القطبية، فمثلاً قد يحتوي المعجم على كلمة «رائع» ولكنها لن تحوي ربما كلمات مثل «رائعة، رائعان، رائعَين، رائعون، رائعِين، إلخ. ولهذا فإن عملية التحليل الصرفي تساعد عمليات التحليل التالية في إدراك أن كلمة مثل «رائعان» مرتبطة بكلمة «رائع» الموجودة في المعجم.

كذلك في اللغة العربية قد تدخل الضمائر على الكلمة، فمثلاً قد يحتوي نص ما على كلمة مثل «حسناتهم»، والتي هي مكونة من قسمين: «حسنات» وهي جمع «حسنة» والضمير «هم». فالمعاجم القطبية قد تحوي كلمة مثل «حسنة» ولكنها لن تحوي الأشكال الصرفية الأخرى أو الحالات التي يدخل فيها ضمير على الكلمة.

- **التجذير والتجذيع Stemming and Lemmatization**

وهما عمليتان تحاولان تجريد الكلمات من الزوائد الصرفية التي تدخل عليها وتحويل الكلمة إلى جذرها الصحيح (كما في التجذير) أو صورة قريبة من الجذر (كما في التجذيع)، ويلجأ الباحثون إلى استخدام هذا النوع من المعالجة مع الطرق المعتمدة على تعلم الآلة بهدف تصغير فضاء المعرفة اللغوية الذي تحتاج الخوارزميات إلى تعلمه حتى تتمكن من تحليل النصوص وتصنيفها.

- **الكشف عن الإشارات المشتركة Co-Reference Resolution**

ويقصد به التعرف على الإشارات المختلفة في النص التي تشير إلى الشيء نفسه سواءً كانت هذه الإشارات على شكل ضمير يعود على الشيء، أو إشارة إلى الشيء باختصار أو جزء من الاسم. فمثلاً في جملة: «أفضل شركة سامسونج على آبل بسبب تجربتي السيئة مع منتجاتها»، الكلمة القطبية «سيئة» موجهة نحو منتجات الجهة المشار إليها بالضمير «ها» الملتصق بالكلمة، وحتى تتمكن تقنيات تحليل الآراء من ربط هذا الرأي القطبي بشكل صحيح يلزم تمييز أن الضمير «ها» هنا يشير إلى شركة آبل كما هو مفهوم من السياق.

- **تصنيف أقسام الكلام Part of Speech Tagging**

ويتم فيها تصنيف كل كلمة في النص بحسب حالتها الصرفية وبحسب سياقها الإعرابي، كتصنيف الكلمة من حيث كونها فعل أو اسم أو حرف، وتمييز الفعل من حيث كونه ماضيا أو مضارعاً أو أمراً، أو تصنيف الاسم على أنه مفرد أو مثنى أو جمع، وتمييز الحروف على أنها أدوات عطف أو وصل أو تأكيد، وتمييز الأسماء إلى صفة أو حال، أو غير ذلك.

وهذه العملية مهمة لحاجة تطبيقات تحليل الآراء إلى التعرف على الصفات. فكثير من الكلمات القطبية صفات، كما أن هذه العملية تسهم في كشف الغموض الذي قد يكتنف بعض الكلمات إذا ما عوملت منفصلةً عن سياقها. مثال لذلك في اللغة العربية كلمة «ذهب» ففي بعض السياقات هي اسم معدن ثمين وتستخدم بشكل متكرر كصفة إيجابية، وفي سياقات أخرى هي فعلٌ ماضٍ للمفرد الغائب.

- **تحليل البناء النحوي Syntactic Parsing الإعراب Dependency Parsing**

تهدف عملية تحليل البناء النحوي إلى كشف بنية الجملة من الناحية النحوية، كتبيان مثلاً أن جملة ما تتكون من شرط وأداة شرط وجواب شرط، أو تحديد الكلمات المكونة لعبارة اسمية Noun Phrase أو عبارة فعلية Verb Phrase.

أما الإعراب فيهدف إلى كشف العلاقات الاعتمادية والمعنوية بين الكلمات، مثل تحديد الفاعل والمفعول به والمفعول لأجله، إلخ.

وكما ذكرنا سابقاً فإن الكثير من طرق تحليل الآراء على مستوى الجمل تحتاج إلى تحليل البناء النحوي والإعراب حتى تتمكن من ربط الكلمات القطبية بمصدرها وبالجهة التي تستهدفها، وتحتاجه كذلك لتعرف إذا كانت التعبيرات القطبية تقع في سياق منفي مثلا بما يستدعي عكس قطبيتها.

والآن، نستعرض طرقاً مختلفة لتحليل الآراء نصنفها إلى:

- طرق تعتمد على خوارزميات مصاغة بشكل يدوي Hand-crafted Rules وتستخدم موارد لغوية كمعاجم قطبية وغيرها.

- وطرق تعتمد على تقنيات تعلم الآلة التقليدية.
- وطرق التعلم العميق.

وهذا التصنيف يمثل أيضاً التطور الزمني الذي مرت به طرق تحليل الآراء، فالطرق المعتمدة على الخوارزميات اليدوية والمعاجم القطبية تمثل المحاولات الأولى لتحليل الآراء وقد عمد إليها الباحثون في ظل ندرة النصوص المقترنة بقطبية معروفة بشكل يمكن استخدامه لتدريب خوارزميات تعلم الآلة، ثم مع توفر مثل هذه البيانات بدأت تبرز الطرق المعتمدة على تعلم الآلة كبديل قوي حل محل الخوارزميات المصاغة بشكل يدوي، ثم مع اتساع نطاق الإنترنت وزخم البيانات الذي شهدته الشبكات الاجتماعية وتوفر كميات مهولة من البينات المصحوبة بآراء معروفة القطبية، برزت تقنيات التعلم العميق وأصبحت هي الآن الخوارزميات الأساسية المستخدمة في تطبيقات تحليل الآراء.

١ - الطرق المعتمدة على المعاجم القطبية Sentiment Lexicons

هذه الطرق تستخدم خوارزميات يتم تطويرها بشكل يدوي وتعتمد على دراية مطورها بالمجال الذي يجري تحليل الآراء فيه، وتحتاج إلى استخدام موارد لغوية كمعاجم القطبية، وقوائم أدوات النفي، أو كلمات تفيد التقوية Intensification أو التضعيف Downtoning، مع إلمام بقواعد اللغة وأنواع العلاقات التي تربط المكونات المختلفة للجمل بهدف الكشف عن نطاق النفي إذا وجد، أو ربط التعبيرات القطبية بمصادرها والجوانب التي تستهدفها في النص. الفكرة العامة لهذه الطرق هي أنها تفحص كل كلمة في النص وتبحث عنها في المعاجم القطبية، وتصنف كل كلمة إلى موجبة أو سالبة أو متعادلة، ويتم تعيين قيمة رقمية لكل من هذه القطبيات فكل كلمة موجبة مثلاً يتم التعبير عنها بقيمة عددية موجبة +١ أو +٢ بحسب شدة القطبية في حال توفر معلومات عن شدة القطبية في المعجم المستخدم- وبالمثل فإن الكلمة السالبة يقابلها رقم سالب -١ أو -٢ ، والكلمات المتعادلة يقابلها الرقم ٠ [٥٣][٥٤][٥٥][٥٦].
تراعي هذه الطرق أيضاً وجود ما يؤثر على اتجاه القطبية أو قوتها من خلال مجموعة من القواعد المصاغة بشكل يدوي، فمثلاً إذا احتوت الجملة على أداة نفي ووقعت الكلمة القطبية في نطاق مسافة معينة -مقاسة بالكلمات- من أداة النفي يتم عكس قطبية الكلمة

والقيمة العددية المرتبطة بها، وكذلك إذا تبعت كلمة قطبية إحدى الكلمات التي تؤثر في شدة قطبيتها يتم زيادة أو تقليل القيمة العددية لقطبيتها وفقاً لذلك [٥٧][٥٨][٥٩]. يلي ذلك تجميع هذه القيم على مستوى الجملة ثم على مستوى النص بكامله، وبذلك تكون القطبية النهائية للنص هي مجموع قطبية الكلمات المكونة له.

المشكلة في هذه الطرق هو اعتمادها على توفر معاجم قطبية ثرية، وتستلزم معرفة قوية باللغة المستعملة في النصوص بشكل عام، وبطبيعة الموضوع الذي يجري تحليل الآراء فيه بشكل خاص، وتحتاج إلى صياغة قواعد خاصة لكل من المواضيع المختلفة، وهو ما يتطلب جهداً كبيراً من الباحثين، فمثلاً القواعد التي تصلح لتحليل التعليقات على المنتجات الإلكترونية لا تصلح بالضرورة لتحليل الآراء في النقاشات التي تتناول مواضيع فكرية. هذا بالإضافة إلى أن هذه الطرق هي الأقل من حيث الدقة في نتائجها، ولذلك انصرف اهتمام الباحثين عنها إلى الطرق المعتمدة على تعلم الآلة.

٢- الطرق المعتمدة على تقنيات تعلم الآلة التقليدية Machine Learning

في هذا النوع من المقاربات يتم الاعتماد على تقنيات تعلم الآلة للتعرف على الأنماط اللغوية المرتبطة بالتعبير عن المشاعر والآراء في النصوص، ويلزم فيها توفر نصوص معروفة القطبية، ويلزم قيام الباحث بتعريف عدد من الخصائص اللغوية Features التي يظن أنها مرتبطة بقطبية النص، وبدلاً من صياغة قواعد ومعادلات يدوية لتصنيف قطبية النص ، تقوم خوارزميات تعلم الآلة باكتشاف العلاقات بين الخصائص التي يعرفها الباحث وقطبية النص وبناء نموذج قادر على تخمين قطبية أي نص جديد بمعلومية خصائصه.

ومن أمثلة الخصائص Features التي حاول الباحثون استخدامها في هذا النوع من تحليل الآراء ما يلي:

- خصائص لفظية Lexical Features: ومن أمثلتها المفردات المتتالية n-grams سواءً من خلال رصد وجود أو غياب كل من هذه المفردات Binary Rep-resentation أو من خلال تعداد تكرار كل منها في النص الواحد Term Frequency (TF) وتكرار ظهورها في النصوص المختلفة Document Fre-

quency (DF)، في هذه الحالة يكون كل n-gram في النص عبارة عن خاصية Feature. هذا يعني أن عدد هذه الخصائص قد يكون كبيراً جداً، وهنا تكون تقنيات كالتجذير والتجذيع والتحليل الصرفي مهمة لاسيما في حالة اللغة العربية لأنها تقلل من عدد هذه الخصائص وتجعل خوارزمية التعلم الآلي أقدر على التعلم.

بعض هذه الخصائص ممكن أن تعتمد على المعاجم، مثل تحديد عدد الكلمات القطبية في الجملة، وتحديد إذا ما كان النص يحتوي على أدوات نفي أو تقوية أو تضعيف، إلخ. وفي هذه الحالة لا يتم تعريف قواعد محددة كما في الطرق اليدوية السابقة وإنما يتم إدخال هذه الخصائص لخوارزمية تعلم الآلة، ويترك للخوارزمية أن تتعلم كيفية الاستفادة من هذه المعلومات لتصنيف القطبية.

- خصائص بنائية Structural Features: وهي خصائص متعلقة بتركيب الجملة والكلمات المكونة لها، ومن أمثلتها طول النص، المسافة بين الكلمات القطبية وأداة النفي إن وجدت، موضع ظهور الكلمات القطبية في النص أو الجملة، إلخ.
- خصائص نحوية Syntactic Features: وهي خصائص تتعلق بالبناء النحوي للجملة والعلاقات الإعرابية التي تربط كلماتها، ومن أمثلتها تصنيف أقسام الكلمات Part-of-Speech، وتفيد هذه الخصائص في جعل عملية تحليل الرأي أكثر إدراكاً للسياق فمثلاً بدلاً من استخدام الكلمة فقط مجردة من سياقها، يصبح بواسطة هذه الخاصية معروفاً إذا ما كانت الكلمة استعملت كصفة أو اسم أو فعل، وإذا ما كانت للمفرد أو المثنى أو الجمع، أو إذا كانت للمذكر أو المؤنث، إلخ.

ومن أمثلة هذه الخصائص أيضاً العلاقات النحوية التي تربط الكلمات مثل ارتباط المبتدأ بالخبر في الجملة الاسمية، والفعل بالفاعل في الجملة الفعلية، إلخ. ومثل هذا الخصائص تكون ضرورية أكثر في حالة الحاجة إلى ربط كل كلمة قطبية بمصدرها وبالجانب الذي تستهدفه، فبدون أن تكون هذه العلاقات النحوية متاحة لخوارزميات تعلم الآلة يكون من الصعب تعلم هذه العلاقات بشكل مباشر من النص.

جُربت العديد من خوارزميات تعلم الآلة لتعلم تصنيف قطبية الآراء، على رأسها خوارزمية التصنيف المعتمدة على مجموعة النقاط الداعمة Support Vector Machines وهي ربما أكثر الخوارزميات استخداماً في هذا المجال وذلك لكفاءتها في التعامل مع أعداد ضخمة من الخصائص، وخوارزمية بييز البدهية Naive Bayes، وخوارزمية التصنيف بحسب أقرب النقاط المجاورة K-NN، والخوارزميات التي تستخدم مجموعات أشجار القرار Tree Ensembles.

٣- الطرق المعتمدة على التعلم العميق Deep Learning

شهدت السنوات العشر الماضية صعودا كبيرا لتقنيات التعلم العميق في العديد من المجالات وحققت نجاحات باهرة في تحليل الصور Image Processing، وإدراك الكلام المنطوق Speech Recognition، ومعالجة اللغات Natural Language Processing. الميزة الأساسية في هذه الطرق أنها تستطيع التعلم بشكل مباشر من البيانات في صورتها الخام وتعفي الباحث من الحاجة إلى تعريف خصائص محددة بشكل يدوي. الصورة الخام للبينات low-level features تكون عبارة عن الكلمات نفسها بتسلسلها في النص أو حتى مجموعات الحروف المتوالية Character n-grams.

تستخدم هذه الطرق أشكالا مختلفة من خوارزميات الشبكات العصبية Neural Networks ، وينصب تركيز الباحثين فيها على بنية نموذج الشبكة Model Architecture، من البنى المستخدمة بشكل متكرر في مجال معالجة اللغات الشبكات العصبية المتكررة Recurrent Neural Networks ومن أمثلتها شبكات الذاكرة قصيرة المدى الطويلة (Long Short Term Memory (LSTM، والشبكات العصبية المبوبة (Gated Recurrent Neural Networks (GRNN، ومن البنى المشهورة أيضاً الشبكات العصبية الالتفافية (Convolutional Neural Networks (CNN في شكلها المطبق على النصوص فضلاً عن الصور، وأخيراً البنى التي شهدت صعودا كبيرا مؤخرا النماذج المنتبهة لنفسها Self-Attention Models ومن أمثلتها خوارزميات Transformer وBERT من شركة جوجل.

ونظراً لأن هذه الخوارزميات تحاول أن تتعلم من البيانات الخام بشكل مباشر فإنها تحتاج إلى كميات كبيرة جداً من البينات حتى تتمكن من اكتشاف العلاقات الاقترانية

بين الكلمات (أو الحروف في بعض الأحيان) وقطبية الآراء. المثير في هذه التقنيات أنها لا تعتمد اعتماداً كاملاً على التعلم من نصوص معروفة القطبية، فبعض مراحل التعلم لا تحتاج سوى نصوص بدون ضرورة لمعرفة تصنيفها Unsupervised Learning، وتهدف هذه المرحلة إلى تعلم تمثيل معنوي للكلمات Word Embedding وهو عبارة عن مجموعة من الأرقام التي يتم تعلمها بشكل آلي لكل كلمة بحيث تصبح هذه الأرقام بمثابة تمثيل رقمي للمعنى الذي تحمله الكلمة و الذي يتم استنباطه من خلال رصد مئات آلاف السياقات التي وردت فيها الكلمة في ملايين النصوص التي يتم تدريب الخوارزمية عليها، ثم يتم استخدام هذه الأرقام للنيابة عن الكلمات في المراحل المتقدمة من تعليم الخوارزمية والتي يلزم فيها استخدام نصوص معروفة القطبية سواءً بشكل كامل Supervised Learning، أو بشكل جزئي أو ضعيف Weak Supervision كأن يفترض أن احتواء النص على وجه تعبيري ضاحك دليل على أن النص يحمل قطبية موجبة.

ومما يميز هذه التقنيات هو سهولة مواءمتها لتصبح قادرة على تحليل الآراء في مجالات مختلفة من خلال تقنيات Transfer Learning، بحيث إذا تم تعليم الخوارزمية على تحليل الآراء في مجال معين مثل مراجعات الأجهزة الإلكترونية، فإنه لا يلزم إعادة تدريب الخوارزمية من الصفر حتى تتمكن من تحليل الآراء الفكرية في الشبكات الاجتماعية مثلاً. وذلك لأن هذه التقنيات تسمح بالإتيان بالنموذج التي تم تعلمه للمجال الأول ثم مواصلة تدريبه على مدونات نصية من المجال الجديد في عملية تسمى أحياناً «مواءمة المجال» Domain Adaptation أو «المعايرة الدقيقة» Fine-Tuning. ومن ميزاتها أيضاً سهولة إجراء التعلم المتزامن للمهام المختلفة Multi-task learning وهو ما يجعل من الممكن تدريب الخوارزمية لتصبح قادرة على إجراء أكثر من مهمة بشكل متزامن مثل تدريب النموذج على تحليل قطبية مراجعات المنتجات، ومراجعات المطاعم، والآراء الفكرية في آن واحد!

وقد أصبحت طرق التعلم العميق الأكثر استخداماً بين الباحثين المهتمين بتحليل الآراء في اللغات المختلفة، والتي لاقت اهتماماً خاصاً بين الباحثين في اللغة العربية وذلك لأن التعقيد الصرفي والنحوي للغة العربية يجعل الاعتماد على الخصائص المعرّفة

يدوياً صعبا جداً وغير عملي. هذا التعقيد من شدته جعل تقنيات التعلم العميق في تحليل الآراء العربية أقل نجاحا منها في اللغة الإنجليزية مثلاً، وقد وجد الباحثون أن إجراء التحليل الصرفي وتقطيع الكلام بناءً على نتيجة هذا التحليل (بل وإجراء عمليات معالجة مثل التجذير والتجذيع) تعتبر خطوات مهمة لتعظيم النجاح الذي تحققه هذه التقنيات في تحليل الآراء العربية.

مصادر وأدوات

نستعرض في هذا القسم مجموعة من الموارد التي نظن أنها مفيدة في مجال تحليل الآراء، ويمكن أن يستفيد منها من يحاول إجراء أبحاث في المجال، أو يحاول أن يبنى أنظمة لتحليل الآراء. سيقتصر العرض هنا على الأدوات الخاصة باللغة العربية.

١. أدوات المعالجة المسبقة للنص:

نستعرض هنا بعض الأدوات التي يمكن استخدامها لتقطيع النص وإجراء عمليات التجذير والتجذيع والتحليل الصرفي، وغيرها.

ومن الأدوات المتاحة لمعالجة النص العربي AMIRA [٦٠] وتضم أدوات لتنفيذ العديد من المهام الأساسية في معالجة اللغة العربية، كالتقطيع Tokenization، وتصنيف أقسام الكلام Part of Speech Tagging، والإعراب السطحي Shallow Parsing.

ومن الأدوات أيضاً MADA [٦١] وتحتوي الباقة على محلل الصرفي وأداة لتقطيع النص وأداة لتسوية النص Orthographic Normalization، وأداة لتحويل النصوص العربية إلى ترميز ASCII وفق طريقة Buckwalter.

ومن أدوات المتاحة لتحليل البناء النحوي للجملة The Stanford Parser [٦٢] وأيضاً Bikel's Parser [٦٣] وكلاهما يدعمان عدة لغات منها اللغة العربية، ويمكن استخدام نفس الأداتين لتصنيف أقسام الكلام كذلك Part-of-speech tagging. ومن الأدوات التي توفر إمكانية الإعراب وإيجاد العلاقات الاعتمادية للباحثين والمطورين TurboParser [٦٤].

٢. معاجم قطبية عربية

نستعرض هنا بعض المعاجم القطبية العربية، ونعرض نوعين من هذه المعاجم. النوع الأول هو المعاجم المعدة بشكل يدوي، والنوع الآخر المعاجم المبنية بشكل آلي أو شبه آلي.

- **المعاجم المعدة يدوياً:**

من أمثلتها معجم ArabSenti [٤١] ويضم ٣,٩٨٢ صفة تم استخراجها من ٤٠٠ مقال من بين المقالات الموجودة في Arabic Tree Bank [٦٥]، وتم تصنيف هذه الصفات إلى إيجابية وسلبية ومتعادلة على يد ثلاثة من متحدثي اللغة العربية.

ومن الأمثلة أيضاً معجم SIFAT [٦٦] وتم بناؤه بطريقة مشابهة ويحتوي على ٣,٣٢٥ صفة.

ومن المعاجم القطبية المتاحة كذلكNileULex [٤٣]، ويتميز باحتوائه على تعبيرات متعددة الكلمات بالإضافة إلى الكلمات المفردة، كما أنه يضمن كلمات وتعبيرات عامية باللهجة المصرية بالإضافة إلى الفصحى، بالمجمل يحتوي المعجم على ٥,٩٥٣ عبارة أو مفردة قطبية. وتتوفر نسخة مطورة من هذا المعجم WeightedNileULex تضيف وزنا يمثل قوة قطبيته [٤٤].

- **المعاجم المعدة بشكل آلي أو شبه آلي:**

ومنها ArSenL [٦٧] ويحتوي على ٢٩ ألف جذر عربي مع أوزان يحدد قوة قطبية كل منها. وArSEL [٦٨] وفيه تم تصنيف الكلمات في المعجم إلى ٨ أنواع من المشاعر مع إعطاء وزن لكل منها.

ومن هذه المعاجم أيضاً SLSA [٦٩] الذي يضم قرابة ٣٥ ألف جذر عربي مع تصنيف قطبية وشدة قطبية كل منها.

٣. مكتبات برمجية:

من أنظمة تحليل الآراء المتاحة للغة العربية نظام SAMAR [٧٠] وهو نظام لتصنيف موضوعية الكلام Subjectivity Analysis وكذلك لتصنيف القطبية Sentiment Analysis. وهو غير متوفر للتحميل عبر الإنترنت ولكن يمكن الحصول عليه بطلبه من أصحاب البحث.

ومن الأنظمة كذلك نظام تحليل المشاعر العربية Arabic Sentiment Analyzer [٧١] وهو متاح للاستخدام عبر الإنترنت ومن خلال المتصفح.

٤. مدونات لغوية Corpora

نستعرض في هذا القسم بعض المدونات اللغوية التي يمكن استخدامها في أبحاث تحليل الآراء العربية، هذه المجموعات تحتوي على نصوص يتم تصنيف قطبيتها بشكل يدوي وفق إرشادات يضعها الباحثون، وتستخدم في طرق تحليل الآراء التي تعتمد على تقنيات تعلم الآلة، كما تستعمل لتقييم قدرة الخوارزميات المختلفة على تصنيف الآراء بشكل صحيح.

المدونة اللغوية المستخدمة في [٢٩] تضم ٢,٨٥٥ جملة تم تصنيف موضوعيتها وقطبيتها بشكل يدوي، ويمكن استخدام هذه المجموعة للدراسات المهتمة بتحليل الموضوعية و/ أو تصنيف القطبية.

المجموعة النصية AWATIF [٧٢] هي امتداد للمجموعة السابقة وفيها أضاف الباحثون ٥,٣٤٢ جملة من صفحات النقاش في ويكيبيديا، و٢,٥٣٢ جملة من منتديات حوار عربية لتيصبح حجم المجموعة ١٠,٧٢٩ جملة تم تصنيف قطبيتها يدوياً.

من المدونات اللغوية أيضاً مجموعة LABR [٧٣] وهي تضم أكثر ٦٣ ألف من تقييمات الكتب مأخوذة من أحد مواقع الكتب، وفيها تعليقات على الكتب كتبها أكثر من ١٦ ألف مستخدم، وكل تعليق مقترن بتقييم رقمي من ١ إلى ٥. مجموعة BRAD [٧٤] هي مجموعة نصية أخرى تحتوي على أكثر من نصف مليون من تقييمات الكتب، وكل التعليقات أيضاً مقترنة بتقييم رقمي من ١ إلى ٥ يدخله صاحب التعليق.

مدونة لغوية أخرى هي HARD ، وهذه المرة تضم هذه المجموعة تقييمات فنادق باللغة العربية يقترب عددها من نصف مليون تقييم مأخوذة من موقع booking.com الشهير، وكما في المجموعات السابقة كل تقييم نصي يأتي مصحوباً بتقييم عددي من ١ إلى ١٠ يدخله صاحب التعليق.

الخلاصة

معالجة الآراء واحدة من أكثر موضوعات لسانيات الحاسب الآلي نشاطاً سواءً في الوسط البحثي أو الوسط العملي، وتطبيقاتها كثيرة ومتشعبة وتلامس جوانب عديدة من حياة الناس. تشتمل معالجة الآراء على مجموعة من المهام الفرعية الأساسية كالتعرف على موضوعية الكلام وقطبيته ومصدره والجهة المستهدفة به، ومهام متقدمة تحتاجها بعض التطبيقات كتلخيص الآراء وتتبع تطورها وكشف انقسام الناس حولها إلى مجموعات. يمكن تصنيف المقاربات التي لجأ إليها الباحثون في هذا المجال إلى ثلاثة أصناف: مقاربات تعتمد على المعاجم القطبية، ومقاربات تعتمد على خوارزميات تعلم الآلة التقليدية، ومقاربات تعتمد على تقنية التعلم العميق الحديثة. تحليل الآراء العربية تواجهه تحديات خاصة نظراً للثراء الصرفي للغة العربية وتعدد لهجاتها، وغياب التشكيل من معظم النصوص المكتوبة بها. ولهذا السبب فإن للمعالجة المسبقة للنص العربي قبل إجراء عمليات تحليل الآراء عليه لها أهمية كبيرة في زيادة دقة تحليل الآراء. ومن هذه المعالجات المفيدة التحليل الصرفي، والتجذيع، والتجذير، والإعراب، وتصنيف أقسام الكلام وغير ذلك. الجهود البحثية في تحليل الآراء العربية أسفرت عن مجموعة غير قليلة من الأبحاث المنشورة والمدونات النصية والمكتبات البرمجية المفيدة في إجراء البحوث وبناء التطبيقات العملية لها.

المراجع

[1] V. S. Poythress، Symphonic theology: The validity of multiple perspectives in theology.، Zondervan، 1987.

[2] P. Heelan، Nietzsches perspectivalism: A hermeneutic philosophy of science، Boston Studies in the Philosophy of Science، 1999.

[3] J. D. Haynes، Perspectival Thinking for Inquiring Organisations.، Informing Science، 2000.

[4] R. Schacht، Making sense of Nietzsche: Reflections timely and untimely، University of، 1995.

[5] L. Doleezel.، Narrative modes in Czech literature.، University of Toronto Press، 1973.

[6] B. A. Uspenskij، A Poetics of Composition: The Structure of the Poetic Text and Typology of a Compositional Form، Univ of California Press، 1973.

[7] C. J. Fillmore، The case for case، UC Berkeley Linguistics، 1967.

[8] M. W. Crocker، Computational psycholinguistics، Department of Computational Linguistics and Phonetics، 2009.

[9] A. Banfield، Unspeakable Sentences: Narration and Representation in the Language of Fiction، Routledge Revivals، 1982.

[10] J. W. Sedelow، Computational sociolinguistics، 1967.

[11] J. Wiebe، Tracking point of view in narrative، Computational Linguistics، 1994.

[12] S. C. Greene، Spin: lexical semantics, transitivity, and the identification of implicit sentiment، ProQuest، 2007.

[13] J. Wiebe، E. Breck، C. Buckley، C. Cardie، P. Davis، B. Fraser، D. Litman، D. Pierce، E. Riloff، T. Wilson، D. Day و M. Maybury ، Recognizing and Organizing Opinions Expressed in the World Press، AAAI Spring Symposium on New Directions in Question Answering، 2003.

[14] L. Zhuang، F. Jing، Zhu و Xiao-Yan، Movie review mining and summarization.، Proceedings of the 15th ACM international conference on Information and knowledge management، 2006.

[15] McDonald، I. Titov و Ryan، A joint model of text and aspect ratings for sentiment summarization، Urbana، 2008.

[16] M. Hu و B. Liu، “Mining and summarizing customer reviews،» تأليف Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining، New York, NY, USA، 2004.

[17] N. Kobayashi، K. Inui و a. Y. Matsumoto، “Extracting aspect-evaluation and aspect-of relations in opinion mining،» تأليف n Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning، 2007.

[18] A. Haenlein و M. K. Michael، Users of the world, unite! The challenges and opportunities of social media، Business Horizons، 2010.

[19] A. Abu-Jbara، B. King، M. Diab و D. R. Radev، “Identifying opinion subgroups in arabic online discussions،» تأليف Proceedings of The Association for Computational Linguistics Conference، 2013.

[20] D. Radev و A. Abu-Jbara، “Subgroup detection in ideological discussions،» تأليف Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics، Jeju, Korea، 2012.

[21] J. Wiebe و S. Somasundaran، “Recognizing stances in online debates،» تأليف Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Conference on Natural Language Processing of the AFNLP، Suntec, Singapore، 2009.

[22] A. Abu-Jbara، J. Ezra و D. Radev، “Purpose and polarity of citation: Towards nlp-based bibliometrics،» تأليف Proceedings of the North American Association for Computational Linguistics، 2013.

[23] R. Jha، A. Abu-Jbara، V. Qazvinian و D. Radev، “NLP Driven Citation Analysis for Scientometrics،» Natural Language Engineering، 2016.

[24] A. Athar و S. Teufel، "Detection of implicit citations for sentiment detection،» تأليف Proceedings of the Workshop on Detecting Structure in Scholarly Discourse، 2012.

[25] S. Teufel و A. Athar، "Context-enhanced citation sentiment detection.،» تأليف Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics، 2012.

[26] A. Athar، "Sentiment analysis of citations using sentence structure-based features،» تأليف Proceedings of the ACL 2011 Student Session، 2011.

[27] M. Abdul-Mageed و M.Diab، "AWATIF: A Multi-Genre Corpus for Modern StandardArabic Subjectivity and Sentiment Analysis،» تأليف Proceedings of the Eight International Con-ference on Language Resources and Evaluation، 2012.

[28] M. Abdul-Mageed و M. Diab، "Toward building a large-scale Arabic sentiment lexicon.،» تأليف Proceedings of the 6th International Global Word-Net Conference، 2012.

[29] M. Abdul-Mageed و M. Diab، "Subjectivity and sentiment annotation of modern standardarabic newswire،» تأليف Proceedings of the 5th Linguistic Annotation Workshop,، 2011.

[30] M. Abdul-Mageed، S. Kuebler و M. Diab، "Samar: A system for subjectivity and senti-ment analysis of arabic social media،» تأليف Proceedings of the 3rd Workshop in ComputationalApproaches to Subjectivity and Sentiment Analysis، 2012.

[31] O. Alharbi، "Classifying Sentiment of Dialectal Arabic Reviews: A Semi-Supervised Approach،» تأليف International Arab Journal of Information Technology، 2019.

[32] H. ElSahar و S. El-Beltagy، A fully automated approach for arabic slang lexicon extraction from microblogs، International Con-

ference on Intelligent Text Processing and Computational Linguistics، 2014.

[33] N. Al-Twairesh، H. Al-Khalifa و A. AlSalman، AraSenTi: large-scale twitter-specific arabic sentiment lexicons، Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics، 2016.

[34] M. Hadzikadic و M. Abdullah، Sentiment analysis on arabic tweets: Challenges to dissecting the Language، Proceedings of the International Conference on Social Computing and Social Media، 2017.

[35] A.-A. e. al، A comprehensive survey of arabic sentiment analysis، Information Processing & Management، 2018.

[36] A. e. al، Survey on Arabic sentiment analysis in twitter، International Science Index، 2015.

[37] A. e. al، A Review on Corpus Annotation for Arabic Sentiment Analysis، International Conference on Social Computing and Social Media، 2017.

[38] A. Assiri، A. Emam و H. Aldossari، Arabic sentiment analysis: A survey، International Journal of Advanced Computer Science and Applications، 2015.

[39] A. Hamdi، K. Shaban و A. Zainal، A Review on Challenging Issues in Arabic Sentiment Analysis، Journal of Computer Science، 2016.

[40] G. BADARO، R. BALY، H. HAJJ، W. EL-HAJJ، K. B. SHABAN، N. HABASH، A. AL-SALLAB و A. HAMDI، A Survey of Opinion Mining in Arabic: A Comprehensive System Perspective Covering Challenges and Advances in Tools, Resources, Models, Applications and Visualizations، ACM Transactions on Asian and Low-Resource Language Information Processing، 2018.

[41] M. Abdul-Mageed، M. Diab و M. Korayem، “Subjectivity and sentiment analysis of modern standard Arabic،» تأليف n Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics، 2011.

[42] M. Karamibekr و A. A. Ghorbani، “Sentence Subjectivity Analysis in Social Domains،» تأليف IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)، 2013.

[43] S. El-Beltagy، “NileULex: A phrase and word level sentiment lexicon for egyptian and modern standard Arabic،» تأليف Proceedings of the International Conference on Language Resources and Evaluation، 2016.

[44] S. El-Beltagy، “WeightedNileULex: A scored Arabic sentiment lexicon for improved sentiment analysis،» تأليف Language Processing, Pattern Recognition and Intelligent Systems. Special Issue on Computational Linguistics, Speech& Image Processing for Arabic Language. World Scientific Publishing Co.، 2017.

[45] Philip، J. Stone و J. Z. N. D. M. O. Robert F. Bales، “The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information،» Computers in Behavioral Science، 1962.

[46] P. Turney و M. Littman، “Measuring praise and criticism: Inference of semantic orientation from association،» ACM Transactions on Information Systems، 21 المجلد، p. 315–346، 2003.

[47] V. Hatzivassiloglou و K. McKeown، “Predicting the semantic orientation of adjectives.،» تأليف EACL، 1997.

[48] J. Kamps، M. Marx، R. Mokken و M. DeRijke، “Using WordNet to measure semantic orientations of adjectives،» تأليف Proceedings of the 4th International Conference on Language Resources and Evaluation، 2004.

[49] A. Hassan، A. Abu-Jbara، W. Lu و D. Radev، “A random walk–based model for identifying semantic orientation،» Computational Linguistics، 3 رقم ،4 المجلد، pp. 539-562، 2014.

[50] A. Hassan، A. Abu-Jbara، R. Jha و D. Radev، “Identifying the semantic orientation of foreign words،» تأليف Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics، 2011.

[51] G. Al-Sukkar، I. Aljarah و H. Alsawalqah، “Enhancing the Arabic Sentiment Analysis Using Different Preprocessing Operators،» تأليف Proceedings of the New Trends in Information Technology، Amman, Jordan، 2017.

[52] A. El-Kholy و N. Habash، “Orthographic and morphological processing for English--Arabic statistical machine translation،» Machine Translation، المجلد vol. 26، pp. pp. 25-45، 2012.

[53] K. Ahmad، D. Cheng و Y. Almas، “Multi-lingual sentiment analysis of financial news streams،» تأليف Proceedings of the 1st International Workshop on Grid Technology for Financial Modeling and Simulation، 2007.

[54] NawafA.Abdulla، NizarA.Ahmed، M. Shehab، M. Al-Ayyoub، M. Al-Kabi و S. Al-rifai، “Towards improving the lexicon-based approach for Arabic sentiment analysis،» Int. J. Inf. Technol. Web Eng، p. 55–71، 2014.

[55] S. Mohammad، F. Bravo-Marquez، M. Salameh و S. Kiritchenko، “Sentiment lexicons for Arabic social media،» تأليف Proceedings of the International Conference on Language Resources and Evaluation، 2018.

[56] H. Awwad و A. Alpkocak، “Performance comparison of different lexicons for sentiment analysis in Arabic،» تأليف Proceedings of the 2016 3rd European Network Intelligence Conference (ENIC’16)، 2016.

[57] M. Elhawary و M. Elfeky، "Mining Arabic business reviews،» تأليف Proceedings of the 2010 IEEE International Conference on Data Mining Workshops (ICDMW'10)، 2010.

[58] R. Duwairi و M. Alshboul، "Negation-aware framework for sentiment analysis in Arabic Reviews،» تأليف Proceedings of the 2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud'15)، 2015.

[59] S. Oraby، Y. El-Sonbaty و M. El-Nasr، "Finding opinion strength using rule-based parsingforArabicsentimentanalysis،» تأليف ProceedingsoftheMexicanInternationalConferenceonArtificialIntelligence، 2013.

[60] M. Diab، "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking،» 2 تأليفnd International Conference on Arabic Language Resources and Tools، 2009.

[61] N. Habash، O. Rambow و R. Roth، "Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization،» تأليف The 2nd International Conference on Arabic Language Resources and Tools (MEDAR)، Cairo, Egypt، 2009.

[62] S. Green و C. Manning، "Better Arabic Parsing: Baselines, Evaluations, and Analysis،» تأليف COLING، 2010.

[63] D. Bikel، "Intricacies of Collins' Parsing Model،» Computational Linguistics، 4 رقم ،30 المجلد، pp. 479-511، 2006.

[64] M. AFT، S. NA و X. EP، "Concise Integer Linear Programming Formulations for Dependency Parsing،» تأليف Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing، Singapore، 2009.

[65] M. Maamouri، A. Bies، T. Buckwalter و W. Mekki، "The penn Arabic treebank: Building a large-scale annotated Arabic corpus،" تأليف Proceedings of the NEMLAR Conference on Arabic Language Resources and Tools، 2004.

[66] M. Abdul-Mageed و M. Diab، "Toward building a large-scale Arabic sentiment lexicon،" تأليف Proceedings of the 6th International Global WordNet Conference، 2012.

[67] G. Badaro، R. Baly، H. Hajj، N. Habash و W. El-Hajj، "A large scale Arabic sentiment lexicon for Arabic opinion mining،" تأليف Proceedings of the Annual Conference on Natural Language Processing، 2014.

[68] G. Badaro، O. El-Jundi، A. Khaddaj، A. Maarouf، R. Kain، H. Hajj و W. El-Hajj، "EMA at SemEval-2018 task 1: Emotion mining for Arabic،" تأليف Proceedings of the 12th International Workshop on Semantic Evaluation، 2018.

[69] R. Eskander و O. Rambow، "SLSA: A sentiment lexicon for standard Arabic.،" تأليف Proceedings of the Conference on Empirical Methods in Natural Language Processing، 2010.

[70] M. Abdul-Mageed، S. Kubler و M. Diab، "SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social،" تأليف Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis، 2012.

[71] M. El-Masri، N. Altrabsheh، H. Mansour و A. Ramsay، "A web-based tool for Arabic sentiment analysis،" تأليف Procedia Computer Science، 2017.

[72] M. Abdul-Mageed و M. Diab، "AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis،" تأليف Proceedings of the International Conference on Language Resources and Evaluation، 2012.

[73] M. Atiya، A. Aly و A. F.، “LABR: A large scale Arabic book reviews dataset.،» تأليف Proceedings of the Annual Meeting of the Association of Computer Linguistics، 2013.

[74] A. Elnagar و O. Einea، “Brad 1.0: Book reviews in arabic dataset،» تأليف Proceedings of the 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA’16). IEEE، 2016.

[75] J. G. M Hernández، “Survey in sentiment, polarity and function analysis of citation،» تأليف Proceedings of the First Workshop on Argumentation Mining، 2014.

[76] H. S. C Jochim، “Improving citation polarity classification with product reviews،» تأليف Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics، 2014.

[77] A. a. H. N. El Kholy، “Orthographic and morphological processing for English--Arabic statistical machine translation،» Machine Translation، 26 المجلد، pp. 25-45، 2012.

[78] M. A.-B. M. D. A. E. K. R. E. N. H. M. P. O. R. a. R. M. R. Arfath Pasha، “Morphological Tagging for Arabic،» [متصل]. Available: http://www1.cs.columbia.edu/~rambow/software-downloads/MADA_Distribution.html. [2019 6 6 تاريخ الوصول].

[79] A. F. a. S. N. A. a. X. E. P. Martins، “Concise integer linear programming formulations for dependency parsing،» تأليف The Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP، 2009.

[80] W. B. a. J. M. Samah Alhazmi، “Arabic SentiWordNet in relation to SentiWordNet 3.0،» تأليف IJCL، 2013.

الباب الرابع

التعلم العميق وتطبيقاته المرتبطة باللغة العربية

د. أحمد الحايك

التعلم العميق وتطبيقاته المرتبطة باللغة العربية

د. أحمد الحايك [1]

ملخص

لقد استطاعت تقنية التعلم العميق (Deep learning) أن تحقق نتائج رائعة في العديد من مجالات الذكاء الاصطناعي وتعلم الآلة خلال الأعوام الأخيرة. يرجع هذا النجاح لعدة أسباب لعل من أهمها توفر وحدات معالجة الرسوميات (GPU) ذات القدرة الحسابية الهائلة وتوفر مجموعات بيانات تدريبية كبيرة جداً تصل إلى ملايين النصوص أو الصور. ويعتبر كل من مجال تحليل النصوص الطبيعية (Natural Language Processing) ومجال تمييز الكلام المنطوق (Speech Recognition) ومجال التعرف الضوئي على الحروف (Optical Character Recognition أو OCR) من أبرز المجالات التي استطاعت تقنية التعلم العميق التفوق فيها على جميع التقنيات التقليدية. هذه المجالات لها أهمية بالغة نظراً لكثرة تطبيقاتها الحالية والمتوقعة، والتي تشمل -على سبيل المثال- التخاطب مع الإنسان الآلي باللغة الطبيعية، والترجمة الآلية. وعلى الرغم من كثرة الإنجازات التي استفادت مؤخراً من تقنية التعلم العميق لخدمة اللغة الإنجليزية وغيرها، إلا أن اللغة العربية لم تستفد بعد من هذه التقنية بشكل كبير.

نقدم في هذا البحث تعريفا لتقنية التعلم العميق وتاريخها وأسباب نجاحها الذي لم يكن يتوقعه معظم الخبراء في مجال الذكاء الاصطناعي. ثم نسلط الضوء بعد ذلك على بعض الأبحاث التي سخرت تقنية التعلم العميق لخدمة اللغة العربية من خلال تطوير خوارزميات عالية الكفاءة في المجالات المذكورة وغيرها، ونأمل أن يكون هذا البحث نقطة انطلاق للاستفادة المثلى من تقنية التعلم العميق لخدمة لغة القرآن العظيم.

١- أستاذ مساعد في كلية علوم الحاسب الآلي بجامعة الأمير مقرن بن عبدالعزيز. حصل د. الحايك على درجة الماجستير من جامعة سارلاند عن خوارزميته لتتميم صور الخلايا ثلاثية الأبعاد، ثم حصل على درجة الدكتوراه في تتبع حركة الإنسان في البيئات غير المنضبطة باستخدام عدد محدود من الكاميرات التقليدية من معهد ماكس بلانك بالتعاون مع جامعة سارلاند. عمل باحثا في معهد ماكس بلانك للمعلوماتية في ألمانيا وباحثا ومدرسا في مركز الأبحاث الألماني للذكاء الاصطناعي في جامعة كايزرسلاوترن، وله العديد من البحوث المنشورة باسمه.

١ - مقدمة

فاز كل من Yann LeCun [١] وGeoffrey Hinton [٢] وBengio Yoshua[٣] مؤخرا بجائزة تورنج (تشبه جائزة نوبل ولكنها تمنح لعلماء الكمبيوتر) لعام ٢٠١٨م [١] بجدارة عن تطويرهم لتقنية التعلم العميق (وتسمى ايضاً الشبكات العصبية العميقة) التي غيرت مسار البحث العلمي في العديد من المجالات وجعلت من بعض الأفكار -التي كان يتصور الكثيرون أنها بعيدة المنال - واقعا نعيشه اليوم. ولأن تطبيقات تقنية التعلم العميق في حياتنا اليومية كثيرة ونتائجها منقطعة النظير، يجدر تقديمها للقارئ العربي.

حتى وقتٍ قريبٍ، كانت الشبكات العصبية الاصطناعية مستبعدة من قبل مجتمع أبحاث الذكاء الاصطناعي. فعلى الرغم من وجودها منذ الأيام الأولى للذكاء الاصطناعي، إلا أنها لم تُنتج سوى القليلِ جداً من النتائج المفيدة عملياً. ولعل أحد أسباب هذا الضعف في الأداء هو أنّ هذه الشبكات مكلفة جداً حسابياً (أي إنها تحتاج إلى إجراء مليارات العمليات الحسابية). بل إن الشبكات العصبية الأبسط منها كانت ربما تحتاج إلى شهور لإتمام عملياتها الحسابية على بعض الحاسبات الآلية الأقدم. بالرغم من هذا، ظلت مجموعةُ من العلماء تبحث في هذه التقنية (مثل Geoffrey Hinton و Yann LeCun اللذان كانا يرأسان مجموعتين بحثيتين لتطوير هذه التقنية [٢]).

قامت مجموعة Geoffrey Hinton بمزامنة هذه الشبكات (أي تقسيمها إلى عدد من المهام التي تنفذ في نفس الوقت على حاسبات آلية متعددة) لإثبات كفاءتها. وفي عام ١٩٩٨م، طورت مجموعة Yann LeCun البحثية مفهوم الشبكات العصبية الالتفافية (Convolutional Neural Network) والتي مكنت من تقليل التكلفة الحسابية للشبكات العصبية وبالتالي زيادة عمقها (راجع الفصل ٣, ٢).

وفي عام ٢٠١٢م، استطاعت تقنية التعلم العميق أن تفرض نفسها بنتائجها الجيدة. فعلى سبيل المثال تمكنت شركة DeepMind التابعة لشركة جوجل من استخدام تقنية

١- رئيس قسم الذكاء الاصطناعي بفيس بوك.

٢- أستاذ فخري بجامعة تورنتو ونائب رئيس شركة قوقل.

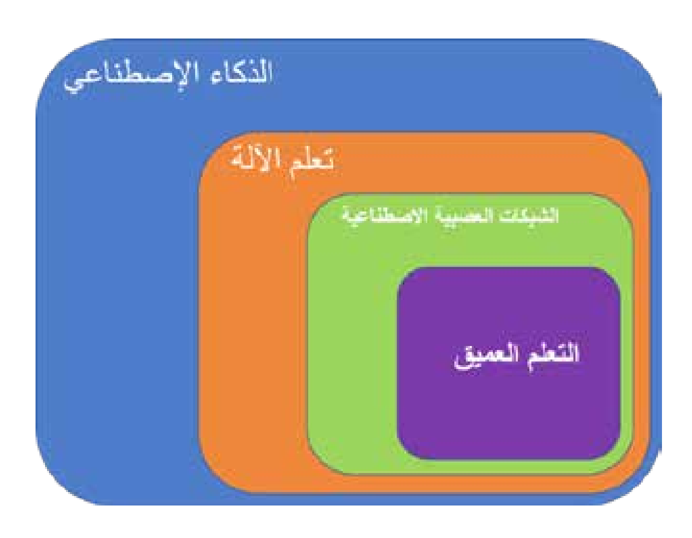
٣- أستاذ بجامعة مونتريال ومدير علمي لعدد من معاهد الذكاء الاصطناعي.

التعلم العميق في تصميم برنامج AlphaGo الذي انتهى به المآل في عام ٢٠١٥م للتفوق على اللاعب الكوري المحترف Lee Se-dol في لعبة Go [٣]. كما تفوقت تقنية التعلم العميق مؤخرا في مجال تشخيص بعض الأمراض كالسرطان رمجال التعرف على الصور (ImageNet challenge) وغيرها من المجالات.

إن فهم فكرة التعلم العميق وأقسامه وتاريخه بشكل تفصيلي يساعد في تسخير هذه التقنية الفعالة. وحتى نفهم المقصود بهذه التقنية، فلابد من تعريف بعض المصطلحات الأساسية مثل: الذكاء الاصطناعي، وتعلُّم الآلة، الشبكات العصبية الاصطناعية؛ لذلك سنفرد الفصل الثاني من الباب للتعرف على معاني هذه المصطلحات قبل أن نسلط الضوء على التعلم العميق وأنواع التقنيات التي استحدثت مؤخرا فيه، كما أننا سنحاول إيضاح أهم أسباب نجاح تقنية التعلم العميق. نعرض بعد ذلك كوكبة من الأبحاث الحديثة التي سخرت هذه التقنية لخدمة اللغة العربية في مجالات تحليل النصوص الطبيعية (Natural language processing)، والتعرف على الكلام المنطوق (Speech recognition)، والتعرف الضوئي على النصوص (Optical Character Recognition)؛ وهي جهود مشجعة نأمل أن تتضاعف حتى نصل إلى تطبيقات ناضجة تخدم اللغة العربية والقرآن الكريم.

٢- تعريف بعض المصطلحات المرتبطة بالتعلم العميق

في هذا الفصل نقدم تعريفات مختصرة لتقنية التعلم العميق وما يرتبط بها من علوم وما يتفرع عنها من التقنيات التي نجحت في تحقيق نتائج قوية خلال الأعوام الأخيرة. ولا شك أن تفاصيل وجوانب التعلم العميق لا يمكن تغطيتها في هذا البحث القصير، لذلك فإننا نعرض في هذا الفصل أفكاره الأساسية دون الخوض في التفاصيل، خاصة وقد أغنت عن الخوض فيها مكتباتٌ برمجية مثل PyTorch [٤] وCaffe [٥] وTensorFlow [٦] التي جعلت بناء خوارزميات التعلم العميق أمرا سهلا وميسراً ووفرت شروحا وأمثلة تيسر ذلك؛ مثل الشروح على عملية التعرف على الأرقام المكتوبة باليد في مجموعة بيانات Mnist [٧].



رسم توضيحي (١): العلاقة بين التعلم العميق والمصطلحات المرتبط به.

ولعل أسهل طريقةٍ لفهم العلاقة بين الذكاء الاصطناعي وتعلُّم الآلة والشبكات العصبية الاصطناعية هو تمثيلها كمجموعات متداخلة كما هو مبين في الرسم التوضيحي ١. فالذكاءُ الاصطناعي هو الدائرة الأكبر؛ إذ خوارزمية تعلم الآلة تعتبر خوارزمية ذكاء اصطناعي والعكس غير صحيح. كما أن تعلم الآلة -بدوره- يشتمل على العديد من الخوارزميات مثل شعاع الدعم الآلي (Support vector machine) والشبكات العصبية الاصطناعية وغيرهما. لذلك فإن الًشبكات العصبية الاصطناعية تعتبر مجموعة جزئية من تعلم الآلة. أما التعلُّم العميق فهو أحد تقنيات الشبكات العصبية الاصطناعية.

١, ٢ الذكاء الاصطناعي

يعرف الذكاء الاصطناعي على أنه علم يهتم بتصميم خوارزميات تستطيع أداء مهام محددة بنفس كفاءة البشر أو أفضل. بناء على هذا التعريف فإن أي خوارزمية تحاكي سلوكا يختص به الإنسان تدخل تحت مظلة الذكاء الاصطناعي. فعلى سبيل المثال، الإنسان يستطيع فهم الكلام، فأي خوارزمية تستطيع عمل هذه المهمة تعتبر خوارزمية ذكاء اصطناعي. وكذلك خوارزميات التعرف على الوجوه في الصور التي تستعمل في برامج Facebook تحملُ بعضَ جوانب الذكاء البشري [٢،٨].

تم اعتماد مصطلح «الذكاء الاصطناعي» عام ١٩٥٦م في مؤتمرات دارتموث [٩]. في ذلك الوقت كان حلم رواد الذكاء الاصطناعي بناءَ آلاتٍ معقدةٍ تمتلكُ حواسا ويمكنها التفكير مثلَ البشر [٢،٨]. وكما أن للبشر قدرة على التعلم مما يسمعونه ويدركونه يشاهدونه، كان تعلم الآلة أحد مجالات الذكاء الاصطناعي التي رمي إلى محاكاة الذكاء البشري عبرها، ومن هنا ظهر مجال «تعلم الآلة».

٢, ٢ تعلم الآلة

تعلم الآلة (Machine Learning) يعنى بتطوير خوارزميات قادرة على تحليل البيانات والتعلم منها لتحسين أدائها في مهمة محددة، كاتخاذ قرارٍ معيَّنٍ أو تصنيف شيءٍ ما. وبعد بناء برامج تعلم الآلة، فإنها تمر بمرحلة تدريب (Training) على بيانات كثيرة مصنفة بشريا لتكسب خوارزمية تعلم الألة القدرةَ على تعلُّمِ تنفيذ نفس المهمة لاحقا على بيانات جديدة غير مصنفة. وهنا، يبدأ الباحثون باختبار أداء الأنظمة (Testing) بعرض بعض المدخلات على الخوارزمية المدربة ومقارنة النتيجة التي تعطيها هذه الخوارزمية بالتصنيف الصحيح لها.

لتوضيح هذا التعريف دعونا نضرب مثالا لخوارزمية لديها القدرة على تحديد نوع الفاكهة التي تظهر في صورة ما. في كل مرحلة، تأخذ هذه الخوارزمية صورة لإحدى الفواكه كمدخل. في المرحلة الأولى يتم بناء الخوارزمية بحيث تكون قادرة على استقبال صور وإعطاء أوسمة محددة كمخرج. ثم تبدأ مرحلة التدريب (Training) بحيث تعطى هذه الخوارزمية عددا كبيرا من صور الفواكه ومع كل صورة تعطى اسم الفاكهة التي تظهر في تلك الصورة، فتقوم الخوارزمية بتحليل كل صورة من أجل إيجاد علاقة بين الصورة ونوع الفاكهة المرفق معها (كالشكل أو اللون أو الحجم) حتى تتمكن الخوارزمية من إيجاد علاقة مطردة بين الصور وأسمائها أو أوسمتها. ثم تبدأ مرحلة الاختبار (Testing) للخوارزمية بأن تعطى بعض الصور الجديدة (أي صور لم تستخدم في مرحلة التدريب) لفواكه من نفس الأنواع التي تم تدريب الخوارزمية عليها؛ ومن ثم، يتم تقييم الخوارزمية وحساب دقتها بتحديد نسبة التصنيفات الصحيحة في مجموعة الصور التي أعدت للاختبار (Testing set).

لقد استطاعت خوارزميات تعلم الالة فتح آفاق واسعة لتطبيقات لم تكن ممكنة بخوارزميات الترميز اليدوي السابقة.كمحركات البحث، وبعض التطبيقات الطبية، والعسكرية، والأمنية، والتجارية، وغيرها [١٠].

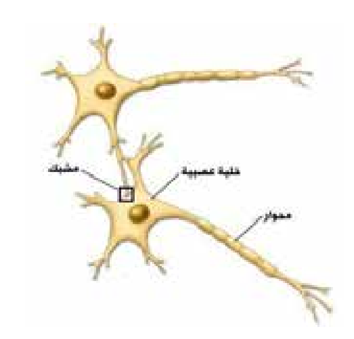
يوجد عدد كبير من خوارزميات تعلم الآلة التي تتبع مناهج مختلفة، مثل: شجرة القرار (Decision tree)، وبرمجة المنطق الاستقرائي (Inductive logic programming)، وخوارزميات المراكمة (Clustering)، والتعلم المعزز (Reinforcement learning)، والشبكات البايزية (Bayesian networks)، وشعاع الدعم الآلي (Support vector machine). ويمكن تصنيف هذه الخوارزميات عموماً إلى مجموعتين رئيسيتين:

- **التعلم تحت الإشراف (Supervised Learning)** وفيه يتم تدريب خوارزمية تعلم الآلة باستخدام بيانات تم وسمها وتصنيفها مسبقاً كما في مثال الفواكه السابق:
- **التعلم دون إشراف (Unsupervised Learning)**: وفيه تجمّع الخوارزمية البيانات المتشابهة إلى مجموعات ومن تطبيقاتها اكتشاف وتصنيف الأشخاص ذوي الاهتمامات المشتركة في وسائل التواصل الاجتماعي [١٠].

ومن بين مناهج تعلم الآلة، ظهرت الشبكات العصبية الاصطناعية لمحاكاة عقل الإنسان في بنيته وطريقة عمله، إذ إن عقل الإنسان يحوي ١٤-١٦ مليار خلية عصبية (أو «عصبونات») مرتبط بعض منها ببعض.

٢,٣ الشبكات العصبية الاصطناعية

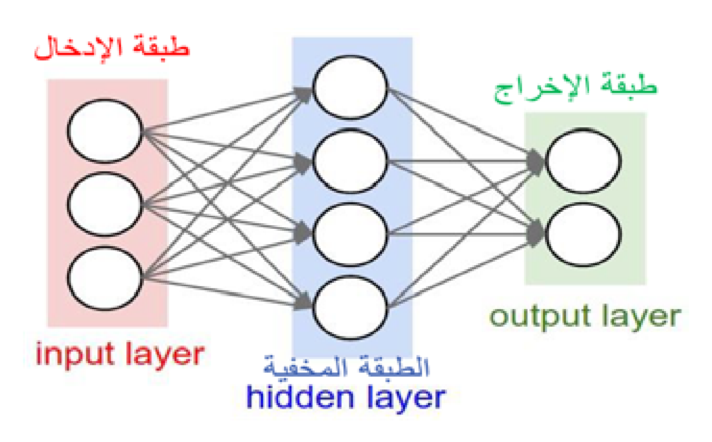
الشبكات العصبية الاصطناعية (Artificial Neural Network. ANN) منهجيةٌ من منهجيات تعلم الآلة مستوحاةٌ من الخلايا العصبية. يوضح الرسم التوضيحي ٢ دماغ الإنسان بشكل مبسط. حيث يُمكن للخلايا العصبية الاتصال بخلايا عصبية مجاورة.





الرسم التوضيحي (٢): الشبكة العصبية الاصطناعية متعددة الطبقات. كل دائرة تمثل خلية عصبية والأسهم تمثل الوصلات بين هذه الخلايا.

وكذلك فإن الشبكات العصبية الاصطناعية تتكون من طبقاتٌ و وصلات لنشر البيانات، أو أوزان مُدخَلاتِ تحسب في مرحلة التدريب ثم يتم تحديدُ الناتج أو التصنيف عبرها أثناء الاستعمال. (أنظر الرسم التوضيحي ٣، حيث تتصل كل خلية بجميع خلايا الطبقة التى تسبقها).



الرسم التوضيحي (٣): الشبكة العصبية الاصطناعية متعددة الطبقات. كل دائرة تمثل خلية عصبية والأسهم تمثل الروابط بين هذه الخلايا.

تتركب الشبكة العصبية الاصطناعية -كما هو موضح في الرسم التوضيحي ٣- من مجموعة من الخلايا العصبية المرتبة ضمن طبقات الإدخال (Input Layer) والإخراج (Output Layer) وطبقة أو أكثر من الطبقات الخفية (Layers Hidden).

وتعتبر الشبكات العصبية بالدخل الأمامي (Feed forward neural network) إحدى أشهر الشبكات العصبية الاصطناعية وقد سميت بهذا الاسم لأنها تعتمد مبدأ الانتشار الأمامي حيث يكون مخرج كل طبقة هو المدخل للطبقة التي تليها فيكون مخرج جميع عصبونات أي طبقة دخلاً لكل عصبون في الطبقة التي تليها. وبزيادة الطبقات الخفية وتطوير خوارزمياتها، ظهر ما يسمى بالتعلم العميق.

٣- التعلم العميق وسر نجاحه

إن مصطلح التعلم العميق (Deep Learning أو DL) اختصار لمصطلح شبكات التعلم العميق (Deep neural networks. DNN)، فإن شبكات التعلم العميق ما هي إلا شبكات عصبية اصطناعية (Neural Networks. NN) ولكنها تحتوى على عدد كبير (أكثر من ١٥٠ طبقة في بعض الحالات) من الطبقات الخفية (Hidden Layers) [٢].

تؤدي هذه الزيادة في الطبقات الخفية لشبكات التعلم العميق إلى زيادة تعقيد عملية التدريب ويتطلب قدرا أكبر من البيانات لتدريبها. وفي مقابل هذه الصعوبة في التدريب فإن الشبكات العصبية العميقة تتميز بالقدرة على تعلم المدخلات بدون الحاجة لتحديد ملامح (Features) مسبقاً خلافا لأكثر خوارزميات تعلم الآلة الأخرى.

تقوم الطبقات الأولى في خوارزميات التعلم العميق تلقائياً بعمليات تنوب عن تحديد وتعلم الملامح بدقة عالية. وبالإضافة لذلك فهي من أفضل الخوارزميات التي تمكن الآلة من تعلم مستويات مختلفة من ملامح البيانات.

فمثلا لو فرضنا أن المدخل للشبكة العميقة صورة، فإن الطبقة الأولى قد تركز على تحديد أماكن الحواف (Edges) في الصورة في حين تركز الطبقة الثانية على تحديد أماكن الزوايا فيها، وهكذا إلى أن تتمكن بعض الطبقات من تحديد الشكل الموجود في الصورة.

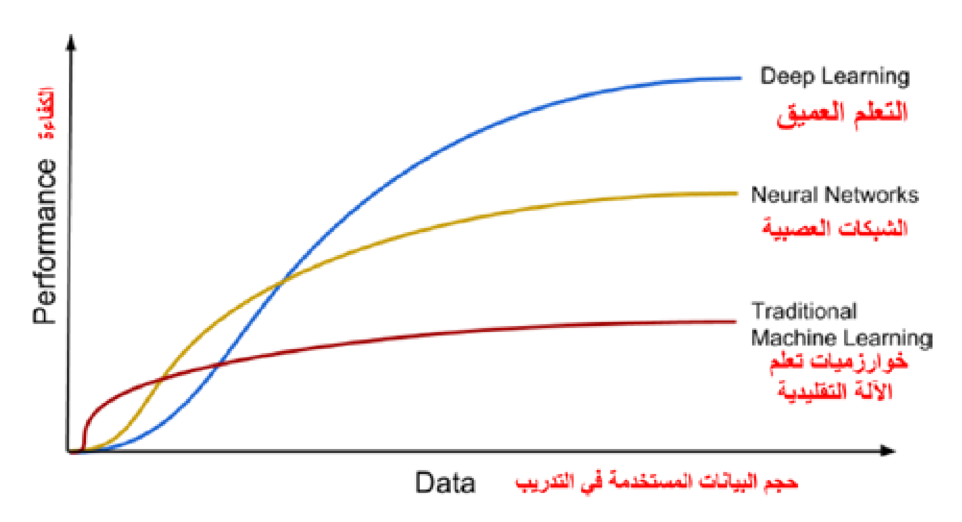
هذا الأمر جعل تصميم أنظمة التعلم بتقنية التعلم العميق أسهل لأنها لا تتطلب الخبرة اللازمة لتحديد ملامح المدخلات، وهو ما قد يعد أهم مراحل خوارزميات تعلم الآلة وأكثرها تأثيراً في نتائجها.

لذا، فإن أحد أهم أسباب نجاح خوارزميات التعلم العميق أنها لا تعتمد على خصائص ثابتة ومحددة مسبقاً كما هو الحال في جميع خوارزميات تعلم الآلة الأخرى، ولكنها تتعلم الخصائص المهمة من البيانات أثناء مرحلة التدريب. غير أن نجاح خوارزميات التعلم العميق يقوم بشكل أساسي على توفر قدر كبير جداً من بيانات التدريب.

يرجع نجاح تقنية التعلم العميق لعدة عوامل منها تطور بعض تقنياتها وخوارزمياتها الحالية مثل الشبكات العصبية الالتفافية (Convolutional Neural Network) التي ساعدت في تقليل التكلفة الحسابية للشبكات العصبية الاصطناعية كثيرا؛ كما سنوضحه قريباً.

كذلك من العوامل التي أسهمت في هذا التطور بشكل كبير توافر وحدات معالجة الرسومات (Graphics processing units) ذات القدرات الحسابية الهائلة، والتي جعلت المعالجَةَ المتوازيَةَ أسرعَ وأرخصَ وأكثرَ قوةً من أي وقتٍ مضى.

من أسباب نجاح التعلم العميق أيضاً توفر كميات كبيرة من البيانات، فقد توفرت مؤخراً كميات هائلة من البيانات وصار بالإمكان جمعها وتخزينها بشكل أسهل وأرخص بكثير من السابق. فهذا التطور الهائل في وحدات التخزين والتدفقِ الهائل للبيانات من كل حدبٍ وصوب، وبكل أنواعها (الصور والنصوص والمعاملات والخرائط... إلخ)، لعب دوراً كبيراً في نجاح تقنية التعلم العميق حيث أن كفاءة التعلم تزداد بشكل مستمر مع زيادة كمية البيانات المستخدمة في مرحلة التدريب. كما يجليه الرسم التوضيحي رقم ٤.



الرسم التوضيحي (٤): مقارنة بين خوارزميات تعلم الآلة من حيث العلاقة بين الكفاءة وكميات بيانات التدريب [١٢].

٤- أبرز تقنيات التعلم العميق

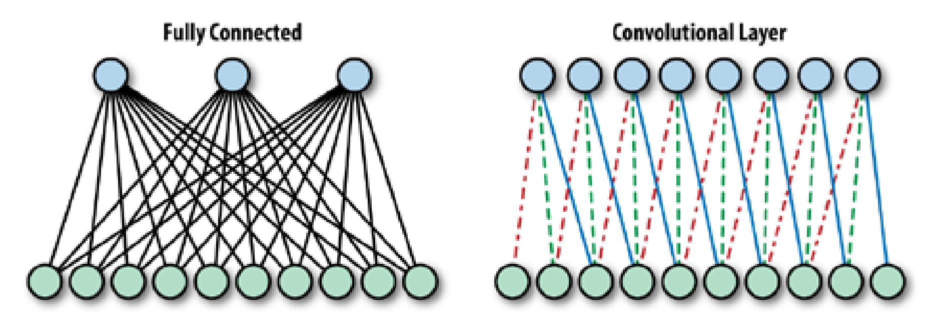
ثمة عدة تقنيات للتعلم العميق، تعتمد على نوع الشبكة العصبية التي تبني منها، وفي هذا الفصل نعرض تقنيات التعلم العميق الحديثة التي حققت نجاحاً كبيراً وانتشاراً واسعاً ونعرج على أسباب نجاحها.

١, ٤ الشبكات العصبية الالتفافية

الشبكات العصبية الالتفافية (Convolutional Neural Network أو CNN اختصاراً) هي نوع خاص وهام من أنواع الشبكات العصبية العميقة قدمها العالم Yann LeCun عام ١٩٩٨م [١٣]. يعتبر هذا النوع من الشبكات العصبية حلاً للكثير من مشاكل الرؤية الحاسوبية (Computer Vision) والتي هي فرع من فروع الذكاء الاصطناعي يعنى بتطبيقات معالجة الصور ومقاطع الفيديو وتحليل محتوياتها.

تقوم الفكرة الأساسية لهذا النوع من الشبكات على استبدال طبقات الاتصال الكامل (Fully Connected Layers) التقليدية بالطبقات الالتفافية (Convolution Layers). ففي هذا النوع من الشبكات تتأثر كل وحدة في الطبقات الالتفافية بعدد محدود من وحدات الطبقة السابقة؛ كما في الرسم التوضيحي ٥.

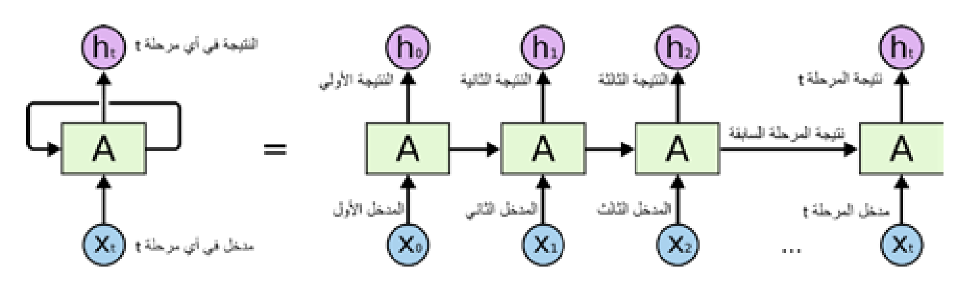
إن فكرة الطبقات الالتفافية مستوحاة من عملية الطي أو الالتفاف الرياضية (Convolution) وهي عملية رياضية تستعمل في تحوير دالة مخرجة من دالتين مدخلتين وتستخدم هذه الأداة الرياضية في الكثير من تطبيقات معالجة الصور. وتقوم طبقة الالتفاف بتطبيق عملية الالتفاف الرياضية على عناصر الدخل (عصبونات الطبقة السابقة أو المدخلات) لحساب قيمة الوحدة في الطبقة التالية.



الرسم التوضيحي (٥): اتصال الطبقة الالتفافية (يمين) والاتصال الكامل (يسار) [١٤].

٢,٤ الشبكة العصبية المتكررة

الشبكات العصبية المتكررة (Recurrent Neural Network أو RNN اختصاراً) من أنواع الشبكات العصبية الاصطناعية والتي تتميز بأنها تتضمَّنُ حلقاتٍ راجعة داخلَ الشبكة ممَّا يُعطيها مفعول الذاكرةِ، فعلى العكس من الشبكات العصبية بالدخل الأمامي (Feed forward Neural Network) فإن الشبكات العصبية المتكررة تأخذ المدخلات على عدة مراحل أو دورات وليس دفعة واحدة ولذلك فإنها تحوي حلقات تعود بالمخرج من الدورة السابقة للخلف بحيث يكون مُدخلا للدورة التالية. هذه الخاصية تعطي الشبكة القدرة على تذكر نتيجة المرحلة الماضية وبالتالي الاستفادة منها في المرحلة التالية. هذه الخاصية مهمة جداً في التطبيقات التي تعتمد على الترابط الزمني بين المدخلات. فعلى سبيل المثال فإن معنى المقطع الصوتي في تطبيقات تحليل الكلام في أي مرحلة يعتمد بشكل كبير على الكلمات السابقة. في مثل هذه التطبيقات تعتبر الشبكات العصبية المتكررة الحل الأمثل. وينبغي التنويه إلى أن تدريبَ الشبكاتِ العصبية المُتكرِّرة مُكلفٌ أكثر من الشبكات العصبونية الالتفافية. الرسم التوضيحي ٦ يبين مخطط الشبكات العصبية المتكررة.



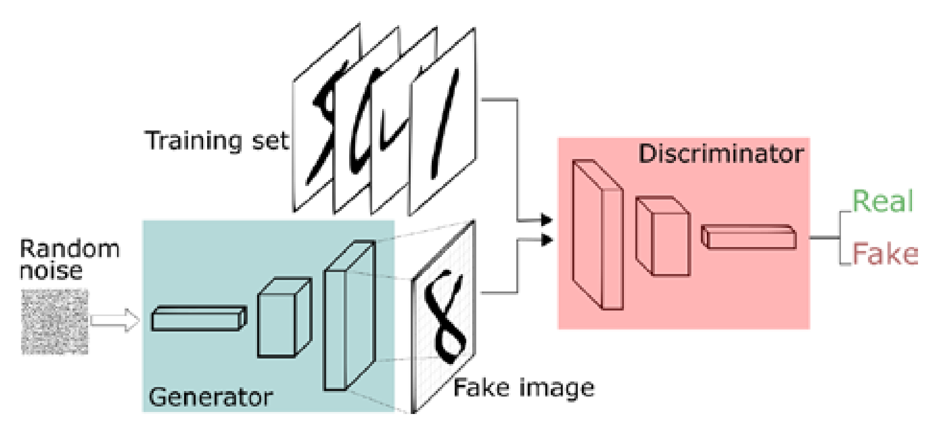
الرسم التوضيحي (٦): تخطيط الشبكات العصبية المتكررة وتمثيل بسطها زمنيا [١٥].

٤,٣ شبكات الذاكرة قصيرة–المدى الطويلة

أحد أهم عيوب الشبكات العصبية المتكررة أنها لا تستطيع التذكر لمدة طويلة. لحل هذه المشكلة تم تطوير شبكات الذاكرة قصيرة–المدى الطويلة (Long Short–Term Memory أو LSTM اختصاراً) نوع خاص من الشبكات العصبية المتكررة RNN مصممة لتخزين نتائج المراحل السابقة لمدد أطول. هذا النوع من الشبكات تمكن من تحقيق نتائج أفضل في الكثير من التطبيقات التي تعتمد على ترابط المدخلات لمدة طويلة [١٦].

٤,٤ شبكات الخصومة التوليدية

شبكات الخصومة التوليدية (Generative Adversarial Networks أو GANs اختصارا) شبكات عصبية عميقة تتألف الواحدة منها من شبكتين متخاصمتين بحيث أن الأولى (وتسمى المولدة (generator)) تسعى لتوليد بيانات تشبه البيانات الحقيقية بشكل كبير، في حين أن الثانية (المميزة (discriminator)) تحاول أن تكتشف إن كانت البيانات المولدة حقيقية أم مزورة، وبعد كل دورة تتعلم كل شبكة وتتطور في مهمتها. فمثلاً، يمكن للشبكة المولدة أن تأخذ صورة لإنسان مرسومة باليد وأن تولد منها صورة معدلة تشبه الأصلية. وعند إدخال الصورة المعدلة للشبكة المميزة، فإن هذه الأخيرة تسعى للحكم على الصورة بأنها حقيقية أو مزورة. ومع التدريب تصبح الشبكة المولدة قادرة على انتاج صورا تشبه الحقيقية إلى حد كبير؛ أنظر إلى الرسم التوضيحي ٧.

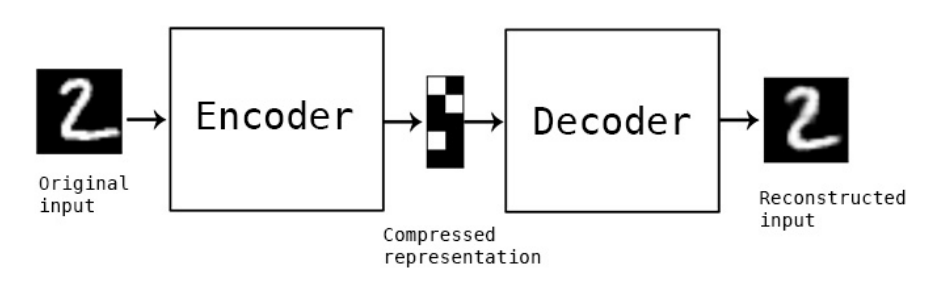


الرسم التوضيحي (٧): مخطط عمل شبكات الخصومةلصور أرقام مكتوبة بخط اليد [١٧].

إن إمكانات شبكات GAN ضخمة، لأنها يمكن أن تتعلم محاكاة أي توزيع للبيانات. وهذا يعني أنه يمكن تعليم الشبكات العصبية إنشاء عوالم تشبه بشكل مخيف عالمنا في أي مجال: الصور، الكلام .تم تقديم GANs في ورقة [١٨] من إعداد Ian Goodfellow وباحثين آخرين في جامعة مونتريال، في عام ٢٠١٤.

٥,٤ شبكة التشفير الآلي

تشبه شبكة التشفير الآلي (Autoencoders) شبكات الخصومة التوليدية حيث إنها تتكون من شبكتين عصبيتين الأولى هي شبكة التشفير (Encoder) وتقوم بتحويل المدخل إلى تمثيل مضغوط (Compressed Representation) والشبكة الثانية هي شبكة فك التشفير (Decoder) وتسعى لإعادة تكوين بيانات الإدخال من خلال تمثيلها الخفي فقط. يتم تدريب كل من هاتين الشبكتين في نفس الوقت بحيث أن الأولى تحاول انتاج تمثيل مخفي يحوي جميع خصائص المدخل مما يمكن شبكة فك التشفير من استرجاع المدخل باستخدام ذلك التمثيل المخفي. بعد انتهاء مرحلة التدريب يفترض أن نصل إلى تمثيل مضغوط يقوم بتمثيل المدخل بشكل دقيق؛ انظر الرسم التوضيحي ٨.



الرسم التوضيحي (٨): توضيح طريقة عمل شبكة التشفير الآلي [١٩].

إحدى فوائد هذه الشبكات أنها تعمل على الحد من حجم المدخلات؛ أي أن حجم التمثيل المضغوط يكون أقل بكثير من حجم البيانات الأولية. فبدلاً من استخدام الصورة ذات الحجم الكبير يمكن استخدام التمثيل المضغوط الذي يقوم مقام هذه الصورة في الكثير من التطبيقات.

٥-أهم تطبيقات التعلم العميق في خدمة اللغة العربية

في هذا الفصل، نعرض بعض تطبيقات تقنية التعلم العميق في خدمة اللغة العربية لتحفيز الجهود في هذا المجال حتى تتطور التطبيقات وتصل إلى مرحلة تمكن من استخدامها في حياتنا اليومية. ورغم أن تطبيقات تقنية التعلم العميق في خدمة اللغة العربية مازلت في مرحلة ابتدائية لم تنضج فيها الخوارزميات المتوفرة لدرجة تؤدي إلى تطبيقات فعالة ومفيدة للمجتمع واللغة، إلا أنه من الصعوبة بمكان استقصاء جميع الجهود التي بذلت في هذا المجال. لذلك فإننا نعرض في هذا الفصل بعض الأبحاث المهمة ذات العلاقة بالموضوع ومن أراد الاستزادة فننصحه بالرجوع إلى بعض الأبحاث الموسعة باللغة الإنجليزية مثل [٢٠].

ومن الجدير بالذكر أن هناك فروق متعددة بين تقنيات تعلم الآلة التقليدية وتقنيات التعلم العميق. من هذه الفروق على سبيل المثال:

- أن تقنية التعلم العميق لا تتطلب خبرة كبيرة في مجال تعلم الآلة على عكس تقنيات تعلم الآلة التقليدية التي تتطلب خبرة كبيرة حيث إنه يقع على عاتق الباحث -في معظم الأحيان- تحويل البيانات الخام إلى ملامح يمكن التقنيات

التقليدية التعامل معها، أما تقنيات التعلم العميق فإنها تتعامل مع البيانات الخام بشكل مباشر دون الحاجة لتحويلها إلى تمثيل آخر. هذا الأمر أدى إلى سهولة استخدام تقنية التعلم العميق.

- نتائج التعلم العميق في خدمة اللغة العربية وغيرها أثبتت تفوقا على تقنيات تعلم الآلة التقليدية. فعلى سبيل المثل في مجال التعرف على الأحرف العربية المكتوبة بخط اليد، استطاعت تقنية التعلم العميق تحقيق نتائج غير مسبوقة.

٥,١ تطبيقات التعلم العميق في مجال تحليل اللغة العربية الطبيعية

تحليل اللغات الطبيعية (Natural Language Processing) هو مجال يعنى بالتفاعلات بين الحاسب الآلي والإنسان من خلال اللغات الطبيعية التي يستخدمها الناس في حياتهم اليومية. في [٢١]، اقترح الباحثون نموذجا لغويا (language model) على مستوى الحرف يقوم بتعيين قيمة محتملة لكل سلسلة من الحروف عن طريق التوزيع الاحتمالي. الجديد في هذا البحث أنه آتى نتائج كانت بالعادة تحتاج لنماذج على مستوى الكلمات. يطبق البحث الشبكات العصبية الالتفافية CNN على أحرف الإدخال قبل إدخالها إلى الشبكات ذات الذاكرة قصيرة-المدى الطويلة LSTM. تم تطبيق هذه الخوارزمية على لغات من ضمنها اللغة العربية. وهذه الخوارزمية متاحة للتنزيل والاستخدام [٢٢].

٥,٢ تطبيقات التعلم العميق في مجال التعرف على الكلام العربي المنطوق

التعرف على الكلام المنطوق (Speech Recognition) هو مجال يعنى بتحويل الكلام المنطوق إلى ترميز حاسوبي نصي. في أحد أول الأعمال التي استخدمت تقنية التعلم العميق في مجال التعرف على الكلام العربي المنطوق [٢٣]، استخدم الباحثون الشبكة العصبية المتكررة للتعرف على الأرقام العربية المنطوقة. تكونت شبكتهم العصبية المقترحة من طبقتين خفيّتين وكان أداؤها جيدًا لبيانات عدة متكلمين.

البحث [٢٤] قدم أفضل حل للتحدي الذي تم اطلاقة عام ٢٠١٧ باسم «Multi-Genre Broadcast» والذي تضمن مجال التعرف على المنطوق. استطاع المؤلفون تحقيق نتائج ممتازة بمزج العديد من التقنيات الحديثة وعلى رأسها تقنية التعلم العميق، حيث

استخدم الباحثون الشبكات العصبية المتكررة مع نماذج لغوية وتقنيات أخرى. كما شارك في تحدي عام ٢٠١٦م باحثون في جامعة لومان بفرنسا وتمكنوا من تسخير تقنية التعلم العميق للتعرف على النماذج الصوتية العربية وتحقق تحسين للدقة بنسبة ٧, ١٥٪ [٢٥].

٣, ٥ تطبيقات التعلم العميق في مجال التعرف على الحروف العربية المكتوبة

يعتبر استخدام تقنية التعلم العميق في مجال التعرف الضوئي (الآلي) على النصوص العربية (Optical Character Recognition أو OCR) من أكثر مجالات خدمة اللغة العربية انتشارا،وإن كان ما زال ثمة مجال للتحسين باستخدام هذه التقنية القوية. في الكتابة، تختلف العربية عن اللغات الأخرى بخصائص منها:

- اتجاه الكتابة في اللغة العربية من اليمين إلى اليسار على عكس اللغات اللاتينية.
- شكل الحرف العربي يعتمد على اتصاله بما حوله.
- طبيعة اللغة العربية مختلفة في الاشتقاق والصرف والنحو والتشكيل وغير ذلك.
- بعض الحروف متشابهة لحد كبير حيث تختلف في بعض الأحيان في عدد أو مواضع النقاط فقط.

هذه الخصائص وغيرها لها تأثير كبير على الأساليب التي يجب أخذها في عين الاعتبار عند دراسة وتصميم خوارزميات التعرف الآلي على النصوص العربية المكتوبة. لذلك فإنه ليس من الممكن دائما تطبيق الخوارزميات المصممة للتعرف على كتابات لاتينية أو صينية -دون تعديلها- على نص عربي.

وبالرغم أن هذا المجال تم بحثه منذ سنوات عديدة من خلال تطبيق تقنيات تعلم الآلة التقليدية إلا أنه لم يبحث بشكل كاف بتقنيات التعلم العميق خصوصاً الجديد من هذه التقنيات. علاوة على ذلك، فإن بعض المشكلات المتعلقة بـمجال التعرف على النصوص العربية لم يتم معالجتها باستخدام تقنية التعلم العميق حتى الآن من هذه المجالات على سبيل المثال التعرف على كاتب النص (Writer Identification) [٢٦].

تقدم [٢٧] نظرة عامة حول مجال التعرف الضوئي على الحرف العربية المكتوبة بخط

اليد. كما أنها تلخص التحديات التقنية الرئيسية المتعلقة بخصائص اللغة العربية. يحاول هذا البحث أيضاً استقصاء البحوث المتعلقة بمجال التعرف الضوئي على الحروف العربية المكتوبة باليد والتى نشرت في عام ٢٠١٥م وما قبله.

في عام ٢٠١٧ قدم الباحث شوقي بوفنار وزملاؤه [٢٨] عملاً استخدم فيه الشبكة العصبية الالتفافية العميقة للتعرف على صور الأحرف العربية المكتوبة بخط اليد. أظهرت نتائج البحث دقة تصل إلى ٣٢,٩٧٪ [٢٩].

وعرض البحث [٣٠] نتائج ممتازة في التعرف على حروف واحدة من مجموعات البيانات المهمة والمعروف باسم (KHATT [٣١]) التي تحتوى على أنماط متنوعة من النص المكتوب بخط اليد، وحقق أداءً متميزاً من خلال تطبيق شبكات الذاكرة قصيرة-المدى الطويلة (LSTM) متعددة الاتجاهات. لقد تمكن باستعمال تقنية التعلم العميق والمعالجة المسبقة من تحسين النتائج من ١٣,٤٦٪ إلى ٨,٧٥٪.

كما قام الباحث أحمد الصاوي وزملاؤه [٣٢] ببناء شبكة عصبية التفافية وتطبيقها للتعرف على الحروف العربية المكتوبة بخط اليد. استخدمت صور وبيانات ١٦٨٠٠ حرف في تدريب واختبار الشبكة لتتمكن من تحيقيق دقة تصل إلى ٩,٩٤٪.

٦- الخاتمة

عرض هذا البحث مقدمة مبسطة للتعريف بتقنية التعلم العميق وأهم ما يرتبط بها من العلوم والمصطلحات، وعرض باختصار عددا من تقنيات التعلم العميق التي حققت شهرة واسعة ونتائج مبهرة. كما سعى للحث عبر أمثلة تطبيقية ناجحة للتقنية على استخدمها في خدمة اللغة العربية. كما يُرجى لهذا البحث أن يكون نقطة انطلاق للتأليف -بالعربية- في مجال التعلم العميق هذا المجال الجدير بالعديد من المؤلفات.

أظهرت تطبيقات التعلم العميق في معالجة اللغة العربية طبيعيا والتعرف على الكلام المنطوق والمكتوب فاعلية رغم من أنها لم تستغل -بعد- بالشكل المرضي. نوصي في ختام هذا البحث بالاهتمام بهذه التقنية التي نتوقع لها نجاحا في الكثير من المجالات وعلى رأسها خدمة اللغات الطبيعية.

المراجع

[1] Association for computing machinery. Fathers of the Deep Learning Revolution Receive ACM A.M. Turing Award. 2018. Retrieved from: https://awards.acm.org/about/2018-turing [Accessed 19 Jun. 2019].

[2] M. Copeland. What's the Difference Between Artificial Intelligence. Machine Learning. and Deep Learning?. Nvidia. 2016.

[3] DeepMind. The story of AlphaGo so far. 2015. Retrieved from: https://deepmind.com/research/alphago/ [Accessed 19 Jun. 2019].

[4] Pytorch. An open source deep learning platform. Retrieved from: https://pytorch.org/. [Accessed 6.6.2019]

[5] Berkeley AI Research. Deep learning framework. Retrieved from: https://caffe.berkeleyvision.org/. [Accessed 6.6.2019]

[6] Tensorflow. An end-to-end open source machine learning platform. Retrieved from: https://www.tensorflow.org/. [Accessed 6.6.2019]

[7] Y. LeCun. C. Cortes. ‚MNIST handwritten digit database‘. (2010) http://yann.lecun.com/exdb/mnist/

[8] H. AlQasir. B. Zeno. W. Dimashky. K. Alsakka. G. S. Saado. H. Azzam. ما هو الفرق بين الذكاء الاصطناعي وتعلم الآلة والتعلم العميق؟ الباحثون السوريون

[9] S. Knapp. Artificial Intelligence: Past. Present. and Future. Vox of Dartmouth. 2006.

[10] F. Al-Qunaieer. "2017 .تعلم الآلة: مقدمة سريعة". https://www.nmthgiat.com.

[11] Memorypsych. The Science of Memory. October 29. 2015. Retrieved from . April 16. 2016. Retrieved from: https://memorypsych.wordpress.com/2016/04/16/the-science-of-memory/

[12] A. Wasicek. Artificial Intelligence vs. Machine Learning vs. Deep Learning: What's the Difference?. sumo logic. 2018

[13] Y. Lecun. L. Bottou. Y. Bengio and P. Haffner. "Gradient-based learning applied to document recognition." in Proceedings of the IEEE. vol. 86. no. 11. pp. 2278-2324. Nov. 1998.

[14] T. Hope. Y. S. Resheff. I. Lieder. Learning Tensorflow: A Guide to Building Deep Learning Systems. O'Reilly Media. 2017.

[15] P. Radhakrishnan. Introduction to Recurrent Neural Network. To Wards Data Science. 2017. https://towardsdatascience.com/introduction-to-recurrent-neural-network-27202c3945f3

[16] F. Gers. Long Short-Term Memory in Recurrent Neural Networks. PhD thesis. 2001

[17] T. Silva. An intuitive introduction to Generative Adversarial Networks (GANs). Free Code Camp.2018. https://medium.freecodecamp.org/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394

[18] I. J. Goodfellow. J. Pouget-Abadie. M. Mirza. B. Xu. D. Warde Farley. S. Ozair. A. C. Courville. Y. Bengio. Generative Adversarial Nets. NIPS (2014).

[19] F. Chollet. Building Autoencoders in Keras. The Keras Blog. 2016 https://blog.keras.io/building-autoencoders-in-keras.html

[20] M. Al-Ayyoub. A. NUSEIR . K. Alsmearat. Deep learning for Arabic NLP: survey. Journal of Computational Science. 2017.

[21] Y. Kim. Y. Jernite. D. Sontag. A.M. Rush. Character-aware neural language models. AAAI (2016) 2741–2749.

[22] Y. Kim. Character-Aware Neural Language Models. github. 2016. https://github.com/yoonkim/lstm-char-cnn

[23] Y.A. Alotaibi. Spoken Arabic digits recognizer using recurrent neural networks. Fourth IEEE International Symposium on Signal Processing and Information Technology. 2004. pp.195–199.

[24] P. Smit. S. R. Gangireddy. S. Enarvi. S. Virpioja and M. Kurimo. Aalto system for the 2017 Arabic multi-genre broadcast challenge. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Okinawa. 2017. pp. 338-345.

[25] N. Tomashenko. K. Vythelingum. A. Rousseau. Y. Estéve. LIUM ASR systems for the 2016 multi-genre broadcast Arabic challenge. IEEE Spoken Language Technology Workshop (SLT). 2016. pp. 285–291.

[26] A. Durou. I. Aref. S. Al-Maadeed. A. Bouridane. E. Benkhelifa. Writeridentification approach based on bag of words with OBI features. Inf.Process. Manag. (2017).

[27] M. Shatnawi. Off-line Handwritten Arabic Character Recognition: A Survey. International Conference on Image Processing. Computer Vision (IPCV). 2015.

[28] C. Boufenar and M. Batouche. Investigation on deep learning for off-line handwritten Arabic Character Recognition using Theano research platform. Intelligent Systems and Computer Vision (ISCV). Fez. 2017. pp. 1-6.

[29] C. Boufenar. M. Batouche. OIHACDB: A New Database for Offline Isolated Handwritten Arabic Character Recognition. COSI. 2016

[30] R. Ahmad. S. Naz. M. Z. Afzal. S. F. Rashid. M. Liwicki. A. Dengel. DeepKHATT: A Deep Learning Benchmark on Arabic Script. Advances in Neural Information Processing Systems. 2017.

[31] S. A. Mahmoud. I. Ahmad. W. G. Al-Khatib. M. Alshayeb. M. T. Parvez. V. Märgner. G. A. Fink. KHATT: an open Arabic offline handwritten text database. Pattern Recognition. 2014.

[32] A. El-Sawy. M. Loey. H. EL-Bakry. Arabic Handwritten Characters Recognition Using Convolutional Neural Network. WSEAS Transactions on Computer Research. 2017.

الباب الخامس

شاعر بلا مشاعر: تجربة في الشعر العربي الآلي باستخدام التعلم العميق

أ. غريب واجب غريبي

شاعر بلا مشاعر: تجربة في الشعر العربي الآلي باستخدام التعلم العميق

أ. غريب واجب غريبي[1]

ملخص

نهدف في بحثنا هذا إلى تسليط الضوء على علم معالجة اللغات الطبيعية Natural Language Processing أو NLP باعتباره أحد أهم مجالات الذكاء الاصطناعي Artificial Intelligence، وسنركز بخاصة على استخدام خوارزميات التعلم العميق Deep Learning فيه لمحاولة محاكاة نصوص الشاعر العربي نزار قباني.

ونستعرض في هذا البحث ماهية علم معالجة اللغات الطبيعية مع إعطاء نبذة تاريخية عن نشأته ومراحل تطوره. ثم نعرج على أهم تطبيقات هذا العلم كالترجمة الآلية، وتحليل المشاعر والآراء، وأنظمة الإجابة التلقائية وغيرها. كما نستقصي أهم الموضوعات الفرعية المندرجة تحت مبحث معالجة اللغات الطبيعية كتصنيف، وتقطيع، وإعراب، وفهم، وتوليد النصوص.

إضافة إلى ذلك، فإننا نشرح أساسيات التعلم العميق وكيفية استخدامه في مجال معالجة اللغات الطبيعية. ومن ثم نستعرض كيفية استخدام الشبكات العصبية المتكررة Recurrent Neural Networks لتوليد نصوص عربية آلياً، حيث نستخدم توليد الشعر العربي الحر كدراسة عملية لهذا الموضوع، فنشرح ذلك -خطوة خطوة- في الفصل الأخير من البحث.

كما أننا نركز على تبيان الأبحاث والأدوات مفتوحة المصدر لمعالجة اللغة العربية عموماً. وذلك من أجل تعريف القارئ بأهم التقنيات والخوارزميات والطرق المستحدثة لجمع، وتمثيل، وتحليل اللغات الطبيعية مع الإشارة إلى بعض المراجع للحصول على معلومات تفصيلية في كل موضوع.

١- باحث مساعد في مركز التعلم الكبير التابع لمؤسسة العلوم الوطنية الأمريكية NSF، ومحاضر في أمن وخصوصية تعلم الآلة، وطالب دكتوراه في جامعة ميزوري بمدينة كانساس الأمريكية UMKC. حصل م. غريبي على درجة الماجستير مع مرتبة الشرف الأولى في تخصص هندسة البرمجيات من جامعة ميزوري بمدينة كانساس، وهو ناشط شغوف في تطوير وإثراء المحتوى العربي للذكاء الاصطناعي.

١ - مقدمة

خلق الله الإنسان وميزه عن باقي مخلوقاته بالعقل، وخلق -سبحانه وتعالى- شعوباً وقبائل لكلٍ منها عاداتها وتقاليدها ولغتها التي تميزها عن غيرها، حيث بينت بعض الدراسات الحديثة [١] أن عدد اللغات الطبيعية (لغات البشر) حول العالم يتخطى الستة آلاف لغة. ومع التطور التقني الهائل في كافة مجالات وعلوم الحاسب الآلي والتقنية، ظهر علم معالجة اللغات الطبيعية الذي يسعى لتمكين الحاسب من فهم ومعالجة وتحليل اللغات الطبيعية لتسهيل الكثير من المهام إلكترونياً في كافة جوانب الحياة.

١,١ التعريف ونبذة تاريخية

علم معالجة اللغات الطبيعية Natural Language Processing أو NLP هو علم تطبيقي يعنى باستخدام تقنيات الحاسب الآلي، وعلى رأسها خوارزميات تعلم الآلة Machine Learning، لأَتمَتَة(١) علوم اللغويات Linguistics بحيث يصبح الحاسوب قادراً على تمثيل وتحليل وتوليد النصوص المكتوبة والمقروءة باللغات الطبيعية كالعربية والإنجليزية.

ومع التطور الهائل في جميع مجالات الحاسب الآلي واستخدامه في أتمتة الكثير من المهام المتكررة، كفهرسة مواقع الإنترنت والرد على العملاء وتحليل آرائهم، أصبح لتقنيات معالجة اللغات الطبيعية دوراً هاماً في الكثير من التطبيقات التي تعتمد على فهم اللغات الطبيعية وتفاعل المستخدمين مع الآلة، كمحركات البحث، والترجمة الآلية، والتلخيص الآلي، وأنظمة الإعلانات الذكية، وتصنيف المواضيع، وتنقية البريد الإلكتروني من الرسائل الضارة، وتحليل المشاعر وقياس الرأي العام، وتوليد النصوص ذات المعنى المترابط والمفهوم.

يمكننا تتبع تاريخ نشأة علم معالجة اللغات الطبيعية إلى خمسينيات القرن الماضي بعد وقت قصير من ظهور الحاسب المعروف باسم Turing Machine [٢] نسبة إلى

---

١ - الأتمتة (Automation): مصطلح مُـعَـرَّب يدل على تحويل العمليات التي تتطلب تدخل البشر إلى عمليات آلية لا تتطلب تدخل البشر. ونعني بها في هذا السياق تطوير برمجيات آلية لا تتطلب تدخل الخبراء لإتمام المهمة.

مصممه العالم الشهير آلن تورينغ، والذي توجه اهتمامه إلى إنشاء برمجيات ذكية تحاكي ذكاء الإنسان. وبالفعل قام في عام ١٩٥٠م باقتراح اختبار تيورنغ Turing Test [٣] للحكم على ذكاء الحواسب من خلال قدرتها على الإجابة بلغة طبيعية على الأسئلة دون قدرة الحكم على تمييز أنها صادرة من حاسب.

ومع تزايد الاهتمام في معالجة اللغات الطبيعية، ظهرت إحدى أولى تطبيقاتها في جامعة جورج تاون لترجمة عبارات بين اللغتين الروسية والإنجليزية [٤-٥]، لتتوالى بعد ذلك تطبيقات كأنظمة إجابة الأسئلة [٦-٧]، وأنظمة تطوير وفهم الحوار [٨-٩]، وأدوات تقطيع الكلام وتحديد أصنافه وإعراب الجمل [١٠-١٣]، وتطبيقات التلخيص الآلي [١٤]، وأنظمة استرجاع البيانات [١٥]، وموخراً ظهرت تطبيقات فهم وتحليل المشاعر والآراء والتي تزامن ظهورها مع انتشار مواقع تقييم المنتجات والخدمات على الشبكة العنكبوتية (الإنترنت) [١٦-١٧].

وكانت أغلب هذه التطبيقات تعتمد على قوانين تصاغ يدوياً من قبل الباحثين ثم تترجم إلى إحدى لغات البرمجة وتعطى للحاسب من أجل تنفيذها. ولكن هذه الطريقة كانت تتطلب فهماً عميقاً للغة وقواعدها ومعانيها بالإضافة إلى الجهد الكبير لتغطية الحالات المختلفة؛ إلى أن ظهرت تقنيات تعلم الآلة في أواسط الثمانينات [١٨-٢٠]، حيث تراجعت الطرق اليدوية السابقة لصالح الطرق الإحصائية التي تترك للحاسب عملية استنباط وتعلم قوانين اللغة بشكل آلي، وذلك من خلال الاطلاع على كميات هائلة من النصوص واستنباط العلاقات المتكررة بينها إحصائياً. وأدت أتمتة هذه الطرق إلى تركيز الباحثين على تحويل النصوص إلى صيغ إحصائية تتمثل فيها أهم خصائص وأنماط اللغة المتكررة. كما وُجدت طرق هجينة تعتمد على الطرق اليدوية لعمل أنظمة خبيرة ومن ثم تضمينها مع تقنيات تعلم الآلة مما أدى إلى تطور تقنيات معالجة اللغات الطبيعية.

ثم نشطت-مؤخراً- خوارزميات التعلم العميق والتي أثبتت قدرتها على معالجة اللغات الطبيعية بشكل يفوق خوارزميات تعلم الآلة السابقة، بما فيها الهجينة، وبدون الحاجة لصياغة النص بشكل إحصائي؛ حيث تعتمد هذه الخوارزميات على بناء شبكات عصبية اصطناعية Artificial Neural Networks يمكنها استنباط القواعد والأنماط

بشكل آلي وبدقة عالية من خلال الاطلاع على كمية كبيرة من النصوص دون الرجوع لقواعد اللغة، كما نبين ذلك في الفصل الثاني.

٢,١ أهم تطبيقات معالجة اللغات الطبيعية

قبل التطرق للوظائف[١] الرئيسية لعلم معالجة اللغات الطبيعية، نسرد في هذا الفصل بعض أهم تطبيقات[٢] معالجة اللغات الطبيعية وبخاصة تلك التي نرى وجوب الاهتمام بها من قبل الباحثين والمبرمجين المهتمين بإثراء معالجة اللغة العربية.

١,٢,١ الترجمة الآلية Machine Translation

لا تخفى أهمية المترجمات الآلية في حياتنا اليومية، إذ هي من أهم -إن لم تكن أهم- تطبيقات معالجة اللغات الطبيعية. وكما ذكرنا في مقدمة الباب، فإن ترجمة النصوص من اللغة الإنجليزية إلى اللغة الروسية كانت أولى خطوات المجال. ومن الأمثلة الأكثر شيوعاً للمترجمات المستخدمة على الإنترنت محرك الترجمة Google Translate من شركة قوقل ومحرك الترجمة Bing من شركة مايكروسوفت. وأول ما بدأت، كانت خوارزميات الترجمة الآلية تتطلب فهماً عميقاً للغات الطبيعية وجهداً كبيراً لتحويلها إلى برمجيات حاسب آلي. وفوق ذلك، فقد كانت دقة وفعالية هذه البرمجيات ضعيفة جداً. ولكن مع انتشار تعلم الآلة -وخاصة التعلم العميق مؤخراً- أصبحت خوارزميات الترجمة الآلية ذات فعالية أكبر وامتدت إلى لغات عديدة، وأصبحت تستفيد من الكم الهائل من النصوص التي يتم إنتاجها بلغات عديدة يومياً على شبكة الإنترنت. وشهدت الترجمة من وإلى اللغة العربية مؤخراً اهتماماً واضحاً كالترجمة للإنجليزية [٢١-٢٥]، وللفرنسية [٢٦-٢٨]. ويمكن الاطلاع على استقصاء للترجمة الآلية من وإلى اللغة العربية في [٢٩-٣٢].

١- الوظائف (Tasks): المهام أو العمليات. فعلى سبيل المثال، عملية إرجاع الكلمة إلى أصلها تعتبر أحد وظائف معالجة اللغات الطبيعية.

٢- التطبيقات (Applications): الاستخدامات. فعلى سبيل المثال، تحليل المشاعر والآراء يعد أحد أهم تطبيقات اللغات الطبيعية.

١,٢,٢ تصنيف النصوص Text Classification

خوارزميات تصنيف النصوص يمكنها الاطلاع على نص معين وتصنيف محتواه إلى موضوعات (كالرياضية، والاقتصادية، والسياسية، وغير ذلك). كثيراً ما يكون تحليل النصوص بناءً على خوارزمية «الورودات الأخيرة» N-grams الشهيرة (المفردة والمزدوجة والثلاثية) والتي تعتمد بشكل عام على تذكر عدد من الكلمات التي تظهر في سياق معين [٣٣-٣٧]. كما تعتمد بعض الخوارزميات الأخرى على استخراج مميزات وخصائص النص [٣٨-٤٠].

حظي هذا المجال ببعض الاهتمام من قبل الباحثين لتصنيف النصوص العربية كاستخدام خوارزميات العد [٤١]، أو تعلم الآلة [٤٢-٤٣]، وكذلك التعرف الآلي (الضوئي) على الحروف [٤٤-٤٥]. ولمن أراد استقصاء الدراسات السابقة لتصنيف النصوص العربية الرجوع إلى المرجع [٤٦].

١,٢,٣ التلخيص الآلي Automatic Summarization

تهتم عملية التلخيص الآلي بتلخيص النصوص، كنشرات الأخبار والتقارير المطولة، واستنباط خلاصتها بشكل آلي. وتساعد عملية التلخيص الآلي في تسهيل كثير من المهام التي تتطلب الاطلاع على خلاصات الكتب والتقارير الطويلة، والبحث عن إجابة معينة داخل النص، واختصار الكلام، وتقليل أحجام الملفات النصية مع الحفاظ على المعاني والمفاهيم الواردة في النص.

وعادة مايتم التلخيص الآلي بإحدى طريقتين: التلخيص الاقتباسي Extractive Summarization [٤٧-٤٨] والتلخيص الخلاصي Abstractive Summarization [٤٩-٥١]. فالتلخيص الاقتباسي يعمل على تلخيص النص من خلال اقتباس أهم العبارات والمفاهيم الواردة فيه بدون توليد أي نصوص جديدة أو اختزال معانٍ غير هامة. وبالتالي فإن جميع الجمل الملخصة هي جمل وتعابير موجودة في النص الأصلي تم تصنيفها من قبل الخوارزمية على أنها مهمة وتلخص الموضوع بقدرٍ كافٍ.

أما التلخيص الخلاصي فيعمل على توليد نصوص تختصر محتوى ومعنى النص الإجمالي باستخدام نص جديد صحيح لغوياً وإملائياً. وبالطبع فإن التلخيص الخلاصي

يحتاج إلى خوارزميات متقدمة تستطيع فهم النص أولاً ومن ثم توليد نص صحيح يلخص النص الأساسي.

وللتلخيص الآلي في اللغة العربية نصيبٌ من الدراسات التي عملت على محاكاة طرق التلخيص في اللغات الأخرى مع الأخذ بعين الاعتبار خصائص اللغة العربية وثرائها النحوي [٥٢-٥٦].

٤ ,٢ ,١ الإجابة على الأسئلة Automatic Question Answering

تعتبر خدمة الإجابة التلقائية على أسئلة وطلبات الزبائن من أنشط المواضيع في مجال معالجة اللغات الطبيعية [٥٧-٥٩]؛ وذلك لأهمية هذا المجال في سوق العمل، وأسواق الأموال، والتجارة الإلكترونية، وغيرها. حيث إن هذه الخوارزميات يمكنها أن تؤدي إلى تطوير برمجيات قادرة على فهم سؤال الزبون، سواءً المكتوب أو المنطوق، ومن ثم البحث عن الإجابة الصحيحة وإيصالها إما نصاً أو نطقاً.

واللافت للنظر في هذا المجال هو جودة وكفاءة عملاء الرد الآلي للغة الإنجليزية حيث يصعب التفريق بينهم وبين العملاء البشر في كثير من الأحيان. ويظهر ذلك جلياً في خدمات الرد الآلي في المتاجر الإلكترونية ومواقع الحكومات الإلكترونية المتطورة.

ومن الأنظمة التي طُورت للرد الآلي باللغة العربية نظام QARAB [٦٠] والذي تم تدريبه على مقالات الصحف العربية وذلك في محاولة لجمع أكبر قدر ممكن من المعلومات عن الأحداث، والتواريخ، والشخصيات وغيرها. وشبيه بهذا النظام نظام AQUSYS [٦١] للرد الآلي على الأسئلة. أما نظام AL-Byan [٦٢] فهو نظام تم تدريبه على نصوص القرآن الكريم للإجابة على الأسئلة الفقهية والموضوعات الدينية.

٥ ,٢ ,١ تحليل المشاعر واكتشاف الآراء Sentiment Analysis

مع التوسع التجاري الهائل في جميع المجالات، وانتشار المنتجات والخدمات المتنوعة على شبكة الإنترنت، ظهرت الحاجة إلى مواقع وخدمات إلكترونية لتقييم المنتجات والخدمات بكافة أنواعها (كالمطاعم، والفنادق، والمدارس، وحتى الدوائر الحكومية). وتتيح هذه الخدمات للمستخدمين كتابة آرائهم وتجربتهم واقتراحاتهم للخدمات الموجودة بحيث يستفيد منها الآخرون بلغة حرة. لذا، كان لابد لصناع القرار ومقدمي

الخدمات التي يتم تقييمها على شبكة الإنترنت من مراجعة هذه التقييمات والمقترحات لتحليلها ودراسة سلوك المستخدمين من أجل تطوير الخدمات وتصحيح أخطائها. وهنا تكمن أهمية خوارزميات تحليل المشاعر والآراء، حيث إنه يصعب على صناع القرار تتبع جميع التقييمات بشكل يدوي على شبكة الإنترنت، وعليه فإن هذه الخوارزميات تلعب دوراً هاماً جداً في تحليل وتلخيص التقييمات بشكل تلقائي وبسرعة فائقة.

تعتمد أغلب أنظمة تحليل الآراء على استباط الكلمات والعبارات ذات دلالات الإعجاب أو الرفض، مثل «المنتج رائع» أو «الخدمة سيئة»، بالإضافة إلى الأخذ بعين الاعتبار الرموز Emojis المستخدمة حالياً في شبكات التواصل الاجتماعي لدلالتها على الإعجاب، أو الحيرة، أو الغضب وغير ذلك.

وبالطبع، فقد اهتم الكثير من الباحثين بتطوير خوارزميات وبرمجيات لتسهيل تجميع وتحليل الآراء باللغة العربية. ومثال ذلك، الدراسة [٦٣] والتي اهتمت بتحليل الآراء وتقسيم مجموعات النقاش على شبكة الإنترنت حسب آراء المشتركين فيها وتوجهاتهم، ونظام SAMAR [٦٤] لتحليل الآراء في شبكات التواصل الاجتماعي باللغة العربية، ونظام [٦٥] لتحليل آراء مستخدمي الفنادق، والدراسة [٦٦] التي سعت لتحليل مشاعر مستخدمي شبكة تويتر للتواصل الاجتماعي.

بالإضافة إلى ذلك، عمل بعض الباحثين على استقصاء أهم الدراسات والأنظمة لتحليل المشاعر والآراء باللغة العربية [٦٧] والتي يمكن الرجوع إليها للمهتمين بتطوير هذا المجال.

٦,٢,١ توليد النصوص Text Generation

عملية توليد النصوص شغلت العديد من الباحثين لأوقات طويلة منذ بدايات ظهور علم معالجة اللغات الطبيعية. وكانت عملية توليد النصوص في بداية الأمر بدائية جداً تعتمد على عمليات الإحصاء والاحتمالات لإعادة توزيع النصوص المدخلة مسبقاً بشكل مختلف [٦٨-٦٩]. وكانت أغلب هذه الطرق تفتقر لوجود ترابط منطقي ودلالي في النصوص التي تم توليدها من قبل الحاسب الآلي.

ومع التطور الأخير في خوارزميات التعلم العميق، وخاصة خوارزميات الشبكات العصبية المتكررة، أصبح مجال توليد النصوص مجالاً خصباً علمياً وعملياً في كثير من التطبيقات. حيث إن توليد النصوص يمكن توظيفة في كتابة المقالات وتلخيص التقارير وعرض النتائج [٧٠-٧٨].

وبسبب ثراء اللغة العربية وقواعدها، يعتبر مجال توليد النصوص العربية أحد أصعب فروع معالجة اللغة العربية. ولكن ومع التطور الحالي في مجالات الذكاء الاصطناعي، وبخاصة التعلم العميق، أصبحت عملية توليد النصوص أقل جهداً بكثير ولا تتطلب تعمقاً في قواعد اللغة بقدر ما تتطلب من خبرات برمجية لبناء خوارزميات لديها القدرة على استنباط قواعد وأنماط اللغة بشكل تلقائي. ونود الإشارة هنا إلى أن مجال توليد النصوص باللغة العربية يعد مجالاً خصباً جداً للدراسة والبحث العلمي وتطوير البرمجيات التطبيقية.

٣,١ أهم وظائف معالجة اللغات الطبيعية

نسلط الضوء في هذا الفصل على أهم وظائف (مهام) معالجة اللغات الطبيعية من أجل تمثيل، وتقطيع، وتجذيع، وربط الدلالات والمعاني في النصوص وغيرها من الوظائف الهامة التي يكثر استخدامها. كما أننا نشير إلى بعض أهم الدراسات والأدوات مفتوحة المصدر التي تهدف إلى خدمة اللغة العربية في هذه المجالات. ونود توجيه القارئ إلى بحث أمجد أبو جبارة «استقصاء تقنيات معالجة اللغات الطبيعية وتطبيقاتها في اللغة العربية» ضمن كتاب «الحرف العربي والتقنية» [٧٩] والذي استقصى فيه الباحث أهم وظائف وتطبيقات علم معالجة اللغة العربية بالتفصيل مع الإشارة إلى أهم مراجع المجال.

١,٣,١ التسوية الإملائية Orthographic Normalization

وتهتم هذه العملية بتجهيز النصوص للمعالجة من خلال إزالة الشوائب الكتابية والرموز التي لا تؤثر في عملية معالجة النص. فقد لا يكون لعلامات الترقيم أو التشكيل أي أهمية في بعض التطبيقات، وعليه يتم إزالتها. ومن الأمثلة الأخرى توحيد الأحرف التي يتم الخلط بينها كهمزات الوصل والقطع، والألف المقصورة والياء في آخر الكلمة، والتاء المربوطة والهاء في آخر الكلمة، وإزالة الـتــطول.

١,٣,٢ التحليل اللفظي Lexical Analysis

ويقصد به تقطيع النص إلى أجزائه الأساسية Tokens من الكلمات والحروف وعلامات الترقيم مع تبيان بداية ونهاية كل وحدة من هذه الأجزاء [٨٠-٨٢]. ونميز هنا بين نوعين للتحليل اللفظي:

(١) التحليل اللفظي السطحي: والذي يعمل على تقطيع النص إلى الوحدات التي تفصل بينها المسافة البيضاء، ونهاية السطر، ونهاية النص، والأرقام، وعلامات الترقيم.

(٢) والتحليل اللفظي العميق: والذي يعمل على تقطيع النص إلى الأجزاء الأساسية للمفردات الناتجة عن تركيب عدة مكونات، كالضمائر المتصلة وأل التعريف.

١,٣,٣ تصنيف أقسام الكلام Part of Speech Tagging

ويُعنى هذا الفرع بدراسة وتصنيف أجزاء الكلام حسب سياقها الإعرابي، كتصنيف الكلمات إلى أسماء (فردية وزوجية وجمع)، أو أفعال (الماضي والحاضر والمستقبل)، أو حروف (كحروف العطف والجر)، وغيرها من أقسام الكلام وتصنيفاته. وتكمن صعوبة هذه العملية في تصنيف أقسام الكلام بناءً على السياق، فيمكن أن تصنف كلمة «سعيد» على كونها اسم أو صفة حسب سياق الكلام.

١,٣,٤ التجذيع Stemming

وهي عملية حذف الزوائد الداخلة على الكلمة لإرجاعها إلى جذعها (أو أقرب ما يكون إلى أساس الكلمة). فكل من المصطلحات التالية arguing وargument و argued تمتلك نفس الجذع argu مع ملاحظة أن هذا الجذع ليس كلمة إنجليزية صحيحة ولكنه الجذع الأقرب لأساس الكلمة. ولعملية التجذيع أهمية في تطبيقات استرجاع البيانات، وفهرستها، وتجميع النصوص، وكشف النصوص المتشابهة [٨٣-٨٦]. ولايزال هذا المجال خصباً للبحث العلمي والتطير في اللغة العربية [٨٧-٨٩].

١,٣,٥ تأصيل الكلمة Lemmatization

وهي عملية إعادة الكلمة إلى أصلها ولكن بشرط كون الأصل كلمة صحيحة وذلك على عكس التجذيع الذي لا يشترط كون الجذع صحيحاً. فأصل كلمة computers

يعود إلى compute (أما جذع الكلمة فهو comput). وثمة اختلاف آخر: إذ إن التأصيل يمكن أن يرجع كلمة إلى أخرى مختلفة في اللفظ كتأصيل am وis وare إلى فعل الكون be.

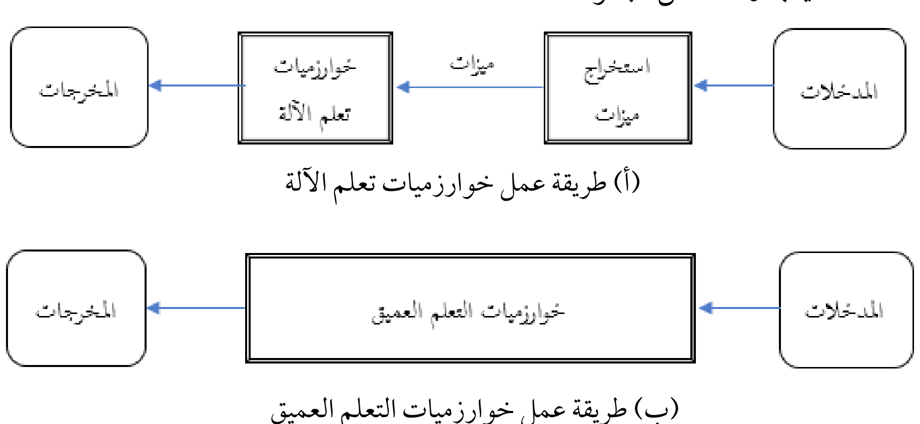
وبالإضافة إلى هذه الوظائف الهامة لمعالجة اللغات الطبيعية، توجد العديد من الوظائف الأخرى التي يستطيع الدارسون الاطلاع عليها، مثل وظائف التشكيل الآلي [٩٠-٩١]، وتحليل البناء النحوي [٩٢-٩٣]، وتحليل علاقات الكلام [٩٤]، وتمييز أسماء الأعلام [٩٥] وغيرها.

ونود الإشارة هنا أنه بالرغم من وجود مصادر متعددة وأدوات مفتوحة المصدر لمعالجة اللغات الطبيعية، إلا أن مجال معالجة اللغة العربية لا يزال يفتقر إلى الكثير من الأبحاث العلمية والعملية والأدوات مفتوحة المصدر للوصول إلى درجات متقدمة تمكننا من تطوير تطبيقات برمجية في مختلف المجالات، وبخاصة تلك التطبيقات التي تعتمد على خوارزميات الذكاء الاصطناعي المتقدمة.

٢-التعلم العميق ومعالجة اللغات الطبيعية

التعلم العميق Deep Learning [٩٦-٩٧] هو أحد فروع علم تعلم الآلة Machine Learning والذي يهتم بتطوير خوارزميات تُمكن الحاسب الآلي من «تعلم» أداء المهام الصعبة التي تتطلب فهماً عميقاً للبيانات وطبيعية عملها (كتشخيص الأمراض تلقائياً باستخدام الصور الطبية). وما يميز خوارزميات التعلم العميق بشكل خاص هو إمكانيتها تعلم المهام بدون برمجة صريحة. ونعني بالبرمجة الصريحة هنا استخراج ميزات البيانات Features بشكل يدوي والحكم عليها بقواعد ثابتة. فخوارزميات التعلم العميق يمكنها استخراج ميزات البيانات وأنماطها المتكررة بشكل تلقائي من خلال الاطلاع على الكثير من البيانات المدخلة ومن ثم تحليلها لإيجاد روابط وعلاقات مباشرة أو غير مباشرة بين البيانات المدخلة (كالصور الطبية) والمخرجات المطلوبة (كتشخيص المرض). وذلك على عكس خوارزميات تعلم الآلة السابقة التي كانت تتطلب فهمَ البيانات وجهداً كبيراً لتحديد ميزاتها وأنماطها بشكل يدوي من قبل علماء البيانات. الشكل ١ يوضح الاختلاف بين طريقة عمل خوارزميات تعلم الآلة

السابقة وخوارزميات التعلم العميق: حيث يوضح الشكل أن خوارزميات تعلم الآلة السابقة تتطلب تدخلاً من قبل علماء البيانات ومختصي المجال من أجل استخراج ميزات البيانات قبل تمريرها إلى خوارزميات تعلم الآلة، أما خوارزميات التعلم العميق فتعمل ذلك تلقائياً بدون تدخل البشر.



(أ) طريقة عمل خوارزميات تعلم الآلة

(ب) طريقة عمل خوارزميات التعلم العميق

الشكل (١): مقارنة بين طريقتي عمل خوارزميات تعلم الآلة والتعلم العميق.

وعلى الرغم من نجاح خوارزميات تعلم الآلة سابقاً في حل الكثير من المشكلات ذات البنى البسيطة، إلا أنها لم تكن فعالة في حل المشكلات ذات البنى المعقدة كاللغات الطبيعة والمشاهد البصرية والإشارات الصوتية. حيث إن هذه المشكلات تتطلب فهماً عميقاً للبيانات وأنماطها وعمل تحويلات غير خطية عديدة ومعقدة من أجل تحويل البيانات بشكلها الطبيعي، كالصورة مثلاً، إلى المخرجات المطلوبة، كوصف محتوى الصورة.

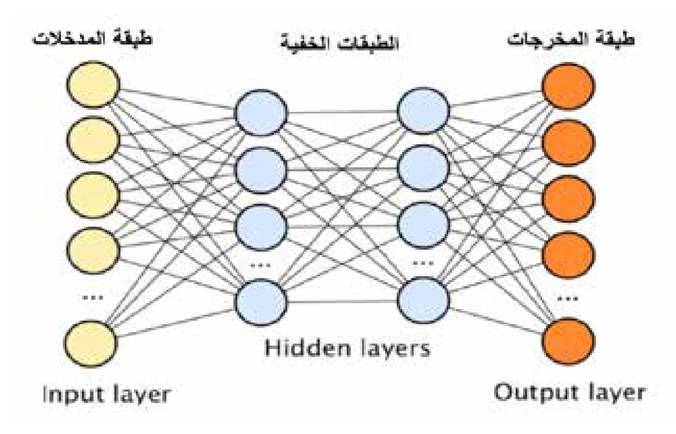
١,٢ كيف تتعلم خوارزميات التعلم العميق

يَكمنُ جوهر خوارزميات التعلم العميق في إيجاد «الرابط العجيب» ما بين البيانات المدخلة كالصور مثلاً، والمخرجات المطلوبة كتصنيف محتوى الصور—راجع الشكل ١ الفقرة ب. وعملية «إيجاد الرابط العجيب» تسمى بعملية تدريب (أو تعليم) الآلة Machine Training. وتتم عملية التدريب من خلال تمرير البيانات المدخلة في عدد كبير من الطبقات المتتالية التي تحوي كل منها على عدد من الوحدات العصبونية (أو العصبونات) Neurons مهمتها تحويل البيانات المدخلة إلى المخرجات المطلوبة من خلال إجراء عمليات رياضية غير خطية عليها.

وتشكل مجموعة الطبقات مايعرف بالشبكات العصبية الاصطناعية Artificial Neural Networks، لكونها مستوحاة من الشبكات العصبية في دماغ الإنسان. وكذلك يطلق عليها مصطلح الشبكات العصبية العميقةDeep Neural Networks بسبب عمق الطبقات فيها (كثرة عددها) وعليه تم تسمية مجموعة خوارزميات تعلم الآلة التي تعتمد على الشبكات العصبية العميقة بالتعلم العميق.

وتختلف خوارزميات التعلم العميق باختلاف بنية Architecture الشبكة العصبية، والتي ترمز إلى عدد الطبقات، وكيفية ارتباطها مع بعضها البعض، وعدد العصبونات في كل طبقة. وبشكل عام، يمكن تصنيف طبقات الشبكات العصبية إلى الأنواع التالية (انظر الشكل ٢):

- طبقة المدخلات: وهي المسؤولة عن إدخال البيانات إلى الشبكة العصبية. وعدد العصبونات في هذه الطبقة مساوٍ لعدد ميزات البيانات المدخلة Features.
- الطبقات الخفية (أو المخفية) Hidden Layers: وتقع مجموعة الطبقات هذه ما بين طبقة المدخلات وطبقة المخرجات، ووظيفتها الأساسية تحويل البيانات المدخلة إلى المخرجات المطلوبة. ويتم تحديد عدد هذه الطبقات وعدد العصبونات داخل كل منها خلال عملية تدريب الشبكة العصبية.
- طبقة المخرجات: وهي المسؤولة عن استقبال نتائج الطبقات الخفية وإصدار النتيجة النهائية للشبكة العصبية (نتيجة التنبؤ Prediction).



الشكل (٢): بنية توضيحية للشبكات العصبية المستخدمة في التعلم العميق (من اليسار إلى اليمين)

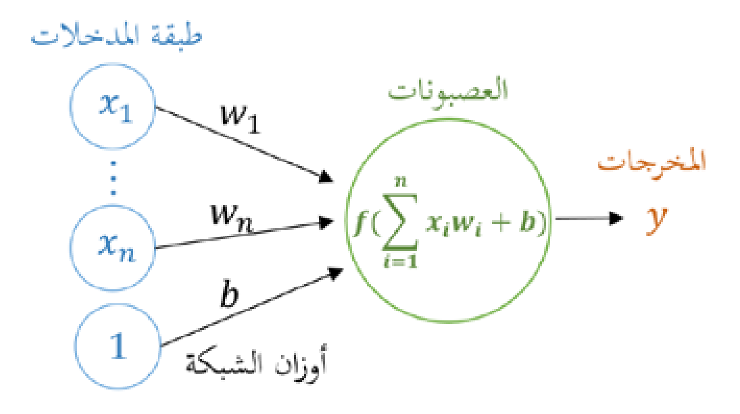
وعملية تدريب الشبكة العصبية—في حالة التدريب تحت الإشراف—تتم بخطوتين أساسيتين: الانتشار الأمامي Forward Propagation و الانتشار الخلفي Back Propagation. تهدف عملية التدريب لضبط أوزان الشبكة (والأوزان هي متغيرات موجودة على روابط الشبكة العصبية تستخدم في حساب نتيجة التنبؤ)، وهي تشمل العمليات التالية في كل من عصبونات الطبقات الخفية، كما في الشكل ٣:

(١) تُضرب كل قيمة من المدخلات $\{x_1 \;\; x_2 \dots x_n\}$ بالأوزان المقابلة $\{w_1 \;\; w_2 \dots w_n\}$،

(٢) وثم تُجمع نتائج عمليات الضرب $\{x_1w_1 + x_2w_2 + \cdots + x_nw_n\}$ وفي بعض الأحيان يتم إضافة قيمة انحياز معينة b للتحكم في نتائج التنبؤ،

(٣) تُطبق عملية غير خطية على نتيجة الجمع من أجل كسر العلاقات الخطية مابين البيانات المدخلة والمخرجات المطلوبة. وتعتبر عملية ReLU إحدى أكثر العمليات الغير خطية المستخدمة في الشبكات العصبية. بعد ذلك، يتم تمرير تلك النتيجة إلى عصبونات الطبقة التالية حيث يتم تكرار هذه العمليات الحسابية في كل وحدة عصبية وهكذا حتى طبقة المخرجات حيث يتم استخراج القيمة النهائية (نتيجة التنبؤ)، وهنا تنتهي عملية الانتشار الأمامي.

(٤) وبعد إيجاد نتيجة التنبؤ يتم مقارنتها مع النتيجة الصحيحة (حيث إننا أثناء عملية التدريب نعرف كل من البيانات المدخلة كالصورة مثلاً ونتيجتها الصحيحة كتصنيف الصورة) بحساب الفرق بين هاتين القيمتين باستخدام دالة خسارة معينة Loss Function، ثم يتم إعادة ضبط أوزان الشبكة بناء على قيمة الخسارة بعملية الانتشار الخلفي Back Propagation من أجل تقليص قيمة الخسارة بأكبر قدر ممكن. ويتم تكرار هاتين الخطوتين (الانتشار الأمامي وثم ضبط أوزان الشبكة) مرات عديدة حتى يتم الحصول على أقل خسارة ممكنة وذلك من خلال إيجاد مجموعة الأوزان المثلى التي يمكن استخدامها لتحويل البيانات المدخلة إلى المخرجات المطلوبة بأكبر دقة ممكنة.



الشكل ٣. المبدأ الأساسي لعمل الشبكات العصبية (من اليسار إلى اليمين).

٢,٢ معالجة اللغات الطبيعية باستخدام التعلم العميق

رغم نجاح الشبكات العصبية في أتمتة الكثير من التطبيقات التي تعتمد على البيانات المنفصلة (كتصنيف الأمراض بالاطلاع على صور الأشعة السينية)؛ فإن هذه الشبكات تواجه تحدياً صعباً عند تحليل البيانات المتسلسلة التي تعتمد على ارتباط وثيق فيما بينها كالنصوص اللغوية والموجات الصوتية ومقاطع الفيديو، حيث إن أتمتة تطبيقات كهذه يتطلب فهماً لسياق النص وتسلسله. لذا، ظهر نوع جديد من الشبكات العصبية التي تملك وظائف إضافية تمكنها من ربط البيانات المتسلسلة حسب ترتيبها الزمني ومن ثم استخدامها في تطبيقات مختلفة كمعالجة اللغات الطبيعية والمشاهد البصرية والأصوات. هذا النوع من الشبكات العصبية، والتي تسمى بالشبكات العصبية المتكررة [٩٨] Recurrent Neural Networks، تمتلك روابط تغذية استرجاعية Feedback Loops تمكنها من اكتشاف الأنماط المعينة ضمن البيانات المتسلسلة الزمنية.

ولكن حتى مع وجود خوارزميات متخصصة في التعلم من البيانات المتسلسلة، إلا أن معالجة اللغات الطبيعية، وبخاصة العربية، لاتزال تواجه العديد من الصعوبات، كتحويل النص إلى ترميز معين يمكن إجراء العمليات الحسابية عليه. ونميز فيما يلي بين طريقتين لتمثيل اللغات الطبيعية في التعلم العميق:

أ) ترميز البت الواحد One-Hot Encoding: وفي هذه الطريقة يتم تمثيل كل كلمة بمصفوفة سطرية (شعاع سطري) تحوي أصفاراً في جميع الخانات ما عدا الخانة التي تمثل

تلك الكلمة (والتي يتم اختيارها بشكل عشوائي غير متكرر) حيث يُوضع الرقم ١ في الخانة المقابلة لتلك الكلمة. فعلى سبيل المثال، نتيجة تمثيل جملة «أكل الطفل التفاحة» قد تتكون من المصفوفات السطرية التالية:

أكل	[١. ٠. ٠]
الطفل	[٠. ١. ٠]
التفاحة	[٠. ٠. ١]

ونلاحظ من المثال السابق أن عدد الأعمدة في المصفوفات السطرية مساوٍ لعدد الكلمات في النص. فلو كان لدينا نص يتألف من ألفي كلمة، لكان حجم تمثيل كل كلمة هو مصفوفة سطرية تحوي ألفي عمود. وهذا بالتأكيد يؤدي إلى إنتاج مصفوفات سطرية ذات حجم ضخم جداً يصعب إجراء العمليات الحسابية عليها، بغض النظر أن أغلب عناصر المصفوفة تحوي أصفاراً.

كما نلاحظ عدم ارتباط معنى الكلمات مع تمثيلها. فعلى سبيل المثال، كلمة «طبيب» يمكن أن تُمثل بـ

[٠ .٠ .٠ .٠ .١]، بينما كلمة «دكتور» يمكن أن تُمثل بـ[١ .٠ .٠ .٠ .٠] رغم احتمال قربهما في المعنى. وهذا بالطبع يفقد تحليل النص أهمية كبيرة في فهم وربط المعاني والجمل والتسلسل المنطقي والدلالات اللفظية وربط الضمائر وغيرها.

ب) تضمين الكلمات Word Embeddings: وهذه الطريقة تعتمد على تمثيل الكلمات باستخدام مصفوفات سطرية مع تضمين العلاقات بين الكلمات المستخدمة [٩٩]. ويتم إنشاء هذه المصفوفات السطرية لتمثيل الكلمات من خلال تدريبها على شبكات عصبية بسيطة البنية. فعلى سبيل المثال، يتم تدريب شبكة عصبية على التنبؤ بالكلمة الناقصة في العبارة التالية «أكل الطفل .... الناضجة». وبالاعتماد على التدريب باستخدام نصوص وفيرة المعاني ذات عبارات مشابهة، فإن كلاً من كلمتي «التفاحة» و»البرتقالة» سوف تمثل احتمالاً عالياً لملئ الفراغ في الجملة السابقة. وهذا يعني أيضاً وجود ارتباط وتشابه بين هاتين الكلمتين (وبالفعل إن الكلمتين متشابهتين في كونهما فواكه). وعليه فإن المصفوفات السطرية التي تمثل كلاً من كلمتي التفاحة والبرتقالة

سوف تحتوي على قيمة رقمية تبين نسبة التشابه والترابط بين الكلمتين. ومن فوائد هذه الطريقة هو تجميع الكلمات ذات المعاني المتشابهة في مجموعات قريبة لبعضها البعض داخل مصفوفات التمثيل. وهذه العلاقات التي يتم تشكيلها بين الكلمات المتشابهة علاقات خطية يمكن تتبعها بسهولة وإجراء العمليات الحسابية عليها. فإذا انطلقنا من مصفوفة التمثيل للمصفوفة السطرية لكلمة «ملك» -مثلاً- ثم تحركنا باتجاه قيمة مشابهة لاتجاه وقيمة المسافة بين كلمتي «رجل» و«امرأة» لوصلنا إلى كلمة «ملكة». وهذا يعني أن كلمتي «ملك» و»رجل» تتواجدان في فضاء رياضي قريب لبعضهما البعض ذات اتجاه موازي لكلمتي «ملكة» و«امرأة». كما أننا إذا طرحنا المصفوفة السطرية لكلمة رجل من كلمة ملك يكون الناتج هو المصفوفة السطرية لكلمة ملكة (ملك - رجل = ملكة).

بعد أن تعرفنا على ماهية التعلم العميق وعلى بعض إمكانيات مجال تحليل ومعالجة اللغات الطبيعية والنصوص المتسلسلة باستخدام الشبكات العصبية المتكررة وعلى بعض طرق تمثيل اللغات الطبيعية، نشرح في الفصل التالي دراسة عملية عن توليد نص شعري عربي حر باستخدام التعلم العميق.

٣- شاعر بلا مشاعر: تجربة في توليد الشعر العربي

نستعرض في هذا الفصل تجربتنا الفريدة في إنشاء الشعر العربي الحر باستخدام خوارزميات التعلم العميق لتوليد النصوص. حيث إننا عملنا على تطوير شاعر إلكتروني، أسميناه «شاعر بلا مشاعر» (لأسباب واضحة) [١٠٠]، يقوم بتوليد نصوص عربية محاكية لأشعار الشاعر الدمشقي نزار قباني (ننشر كثيراً منها كتغريدات في شبكة التواصل الاجتماعي «تويتر»).

نهدف في هذا الفصل إلى تعريف القارئ بالخطوات والمهام اللازمة لتكرار وتطوير هذه الدراسة (كما أننا نوفر المصدر المفتوح لهذا الخوارزمية على الرابط التالي [١٠١]) ونشجع على تطوير أدوات أخرى تعمل على توليد النصوص العربية في مجالات مختلفة، وذلك أن الخوارزمية المستخدمة مفتوحة المصدر (Open Source) ويمكن إعادة استخدامها مجاناً في أكثر من مجال كتوليد الروايات، أو المواضيع التقنية، أو حتى تطوير أنظمة للرد التلقائي على رسائل البريد الإلكتروني.

١,٣ تجميع وتهيئة البيانات

عملية تجميع وتجهيز البيانات واحدة من أصعب وأطول مراحل بناء نماذج تعلم الآلة، إذ يصعب الوصول إلى بيانات جيدة ومفتوحة المصدر لاستخدامها في تدريب هذه الخوارزميات. كما أنه -حتى مع وجود بيانات مفتوحة المصدر- لا بد من بذل الوقت والجهد في تهيئة البيانات لتكون صالحة للاستخدام من قبل خوارزميات التعلم العميق، كعمليات تنظيف البيانات، وتمثيلها (تحويلها من نصوص إلى أرقام)، وتعبئة البيانات الناقصة، وغيرها من الخطوات اللازمة قبل البدء في عملية التدريب.

وقمنا بتجميع البيانات اللازمة (أشعار نزار قباني) في تجربتنا هذه بالطريقتين التاليتين:

- من خلال استخدام محرك البحث جوجل. واعتمدنا هنا على البحث عن مواقع تحوي أشعاراً لنزار قباني ثم قمنا بنسخ ولصق هذا الأشعار داخل ملفات نصية -بعد التأكد من صلاحية حقوق النشر لهذه الأبيات الشعرية.
- من خلال استخدام شبكة توتير للتواصل الاجتماعي. حيث عملنا على تطوير برنامج بلغة «بايثون» ليقوم بالبحث التلقائي عن تغريدات شعرية لنزار قباني [١٠٢] وتحميلها في الملف النصي.

وبعد أن تكونت لدينا مجموعة مناسبة من النصوص (الأبيات الشعرية) لعملية التدريب، عملنا على استخدام وظائف معالجة اللغات الطبيعية التي ذكرناها في الفصل السابق للتنسيق، والتسوية الإملائية، وإزالة الشوائب من النصوص. وبشكل خاص، عملنا على التأكد من خلو النصوص من الكلمات الإنجليزية، والرموز التعبيرية، والدوال التصنيفية (Hashtag) وذلك لعدم أهمية هذه الأجزاء في تدريب الآلة وإنما تعتبر شوائب يجب إزالتها. ونوضح عمليات تجيهز النص في الخوارزمية١.

```
Algorithm: PreprocessText(text)
1:    words = split_text_by_space(text)
2:    for word in words:
3:      if word.startWith('#') || word.isEnglish() || word.isEmoji():
4:         remove word
5:      end if
6:      word.removeExtras()        // إزالة الشوائب النصية كالتطويـــل
7:    end for
8:    return words
```

الخوارزمية ١. توضيح مبسط لخوارزمية تجهيز النص

وبعد إزالة الشوائب من النص، كان لا بد من إيجاد طريقة مناسبة لتمثيل النص. وعلى الرغم من وجود العديد من الطرق لتمثيل النصوص، كما شرحنا سابقاً، إلا أننا اعتمدنا في تجربتنا هذه على تحويل كل حرف ورمز من النص إلى رقم عشري محدد لتسهيل عملية التدريب. وقمنا بعمل ذلك من خلال إنشاء شعاع (مصفوفة سطرية) من الحروف والرموز الفريدة في النص وإعطاء كل منها رقم معين عشوائي بحسب أول ظهور له في النصوص، وبلغ طول الشعاع ٤١ للحروف وعلامات الترقيم والتشكيل.

ولتسريع عملية التدريب، قمنا بتحويل هذه الأرقام إلى أرقام كسرية ما بين الصفر والواحد، وذلك لأن عملية التعلم تتم من خلال ضرب هذه الأرقام بأوزان الشبكة ومن ثم تطبق التحويلات الغير خطية عليها (راجع الشكل ٣). وتحويل الأرقام العشرية إلى كسرية يصغر قيم النتائج فيسرع عمليات الضرب وبالتالي يقلص الوقت اللازم لتدريب الشبكة العصبية.

٢,٣ اختيار وحدة النموذج

بعد تجهيز البيانات، واجهنا الحاجة للاختيار بين طريقتين مختلفتين لتدريب النموذج: إما تدريب النموذج ككلمات متتالية أو كحروف متتالية. فتدريب النموذج على الكلمات -بدلاً من الحروف- يتفوق في توليد نصوص ذات معنى مترابط، حيث إن النصوص المولدة ستحوي كلمات صحيحة دائماً، كما أن وقت التدريب أقل بكثير مقارنةً بتدريب النموذج على الحروف؛ ذلك لأن تدريب النموذج على الكلمات يعني

أن النموذج على دراية سابقة بالكلمات وإنما يهدف لاستنباط سياق الكلام وقواعده وكيفية توزيع الكلمات.

أما تدريب النموذج على الحروف فيحتاج لوقت أطول ولشبكات عصبية ذات بنى عميقة جداً وذلك لأن الشبكة العصبية تحتاج لتعلم إنشاء الكلمات من الحروف والقواعد الإملائية أولاً قبل تعلم استنباط سياق الكلام وكيفة توزيع الكلمات. ولقد اخترنا توليد النصوص حرفاً حرفاً في تجربتنا هذه لسببين أساسين:

- أننا أردنا أن نختبر إمكانية تدريب النموذج على عملية توليد النصوص مع علامات التشكيل. حيث أن بعض الأشعار التي استخدمناها في عملية التدريب كانت مشكلة. وبالتالي فإن عملية تدريب النموذج حرفاً حرفاً سوف تضمن تدريب النموذج على علامات التشكيل باعتبارها حروفاً.
- أننا أردنا–فعلاً– إبراز قدرة الشبكات العصبية على تعلم توليد كلمات عربية صحيحة ذات معنى ودلالات مترابطة من الحروف، بدلاً من إعادة إنشاء كلمات موجودة مسبقاً داخل النص المستخدم في عملية التدريب.

٣,٣ تدريب النموذج

الخطوة التالية تمثلت في تقسيم النص إلى أقسام متسلسلة موحدة الطول لتغذيتها في نموذج التعلم العميق، حيث قررنا استخدام سلاسل نصية مكونة من ١٠٠ حرف لتغذي النظام بشكل دوري أثناء عملية التدريب (وذلك لأن هدفنا كان إنشاء شاعر آلي يقوم بتغريد الأشعار –أو مايشابه الأشعار– على شبكة تويتر ، كما أن هذا الطول مناسب لتدريب الشبكات العصبية بناءً على الحروف عموماً). وأخيراً، قمنا بتمثيل النص بطريقة One-Hot Encoding التي شرحناها سابقاً.

ولتوضيح عملية التدريب، فإننا نزود النظام بمئة حرف في كل دورة وندع له التنبؤ بالحرف التالي حتى يتم تدريب الخوارزمية على كافة النص، ونقوم بتكرار هذا العملية على النص كاملاً مرات عديدة حتى تزداد كفاءة التنبؤ في النموذج. وعليه يمكننا اعتبار عملية تدريب الشبكات العصبية على توليد النصوص بأنها عملية تدريب النموذج على التنبؤ بالحرف التالي في سلسلة نص معينة.

فعلى سبيل المثال، إذا عملنا على تقسم النص إلى متسلسلة ذات طول أربعة حروف في العبارة التالية «سبحان الله»، فإن خطوتي التدريب والتنبؤ سوف تعملان على الشكل التالي:

خطوة التدريب	خطوة التنبؤ
س ب ح ا	ن
ب ح ا ن	(مسافة)
ح ا ن (مسافة)	ا
ا ن (مسافة) ا	ل
ن (مسافة) ا ل	ل
(مسافة) ا ل ل	ه

٤, ٣ اختيار بنية النموذج

من أجل توليد النصوص باستخدام التعلم العميق، يمكن استخدام الشبكات العصبية المتكررة Recurrent Neural Network وبشكل خاص، بنية Long Short-Term Memory لفعاليتها المعروفة في تحليل البيانات المتسلسلة. واعتمدنا في إنشاء شبكتنا العصبية على البنية التالية:

- طبقة المدخلات: وتحوي ١٠٠ وحدة عصبونية مهمتها إدخال السلاسل النصية التي قمنها بتجهيزها سابقاً ومن ثم تمريرها إلى الطبقة الخفية الأولى.
- طبقتان خفيتان: الأولى تحوي ٢٥٦ وحدة عصبونية، والثانية تحوي ١٢٨ وحدة عصبونية (نصف الأولى).
- طبقة المخرجات والتي تستقبل بيانات الطبقات الخفية السابقة وتحولها إلى حرف معين والذي يمثل نتيجة التنبؤ بالحرف التالي للمئة حرف المدخلة في الشبكة العصبية.

واختيارنا لهذه البنية كان بعد العديد من التجارب، حيث لا توجد -حتى الآن- طريقة علمية معتمدة لاختيار البُنى المثلى لشبكات التعلم العميق بسبب عدم معرفة كيفية توزيع الأوزان داخل الطبقات الخفية كما ذكرنا سابقاً. فعملية إيجاد البنية المثلى

للشبكات العصبية (عدد الطبقات الخفية والعصبونات في كل منها) هي عملية بحث تتم من خلال المحاولة والتكرار ومراقبة الأخطاء والتعلم منها.

٥, ٣ تدريب وتقييم النموذج

بعد تطوير بنية نموذج التعلم العميق وتجهيز النص لاستخدامه في عملية التدريب، قمنا بالبدء الفعلي بعملية تدريب النموذج على توليد النصوص. حيث بدأت الشبكة العصبية بتوليد نصوص مقروءة بعد الكَرّة (أو الدورة) Epoch العشرين (والكَرّة هي عملية التدريب الواحدة على كافة النص الموجود). وأكملنا عملية التدريب حتى الدورة الخمسين حيث بدأت الشبكة العصبية بتوليد نصوص ذات نتائج عالية الدقة وصلت حتى ٪٩٣. وهي نتيجة مقبولة جداً لو أخذنا بساطة البنية المستخدمة وعمليتي التدريب وتهيئة البيانات بالإضافة إلى حجم نص التدريب لدينا بعين الاعتبار مقارنة بطرق توليد النصوص التقليدية.

٤ - النتائج

نوضح في الشكل ٤ أمثلة من التغريدات الشعرية التي تم توليدها ونشرها بواسطة شاعر بلا مشاعر. ونلاحظ من خلال هذه الأمثلة أن بعض الكلمات تحوي تشكيلاً، وذلك لأن بيانات التدريب كانت تحوي التشكيل أيضاً. كما نلاحظ أن معظم التغريدات تحوي شطراً شعرياً واحداً وذلك لأننا قمنا بتدريب النموذج على متسلسلات نصية بطول ١٠٠ حرف. ويمكن تعديل ذلك بكل سهولة لإنشاء الشطور الشعرية بأشكال مختلفة، ولكننا اقتصرنا على الشعر الحر هنا للسهولة ومناسبة منصة التواصل الاجتماعي ومحدودية عدد الأحرف فيها. كما نلاحظ أن بعض التغريدات احتوت على كلمات غير مناسبة أو لا معنى لها (ككلمة «المِسرا» في التغريدة الأخيرة في الشكل ٤ مثلا)، وهذا متوقع حيث إن النموذج تم تدريبه على الحروف لا الكلمات.



الشكل (٤): بعض الأمثلة للنصوص التي تم توليدها

وقد لاقت هذه الأداة إعجاب بعض مستخدمي موقع التواصل تويتر وحصلت على ١٠٣ متابع، بمعدل ٣ إعجابات لكل تغريدة حتى تاريخ كتابة هذا البحث. وندعو المهتم إلى الاطلاع على هذه الأداة [١٠٠] تحت اسم المستخدم @AI_Sha3er وإنشاء أدوات مشابهة كتوليد القصص والروايات.

٥- الخاتمة

قدمنا في بحثنا هذا لمحة مبسطة عن معالجة اللغات الطبيعية، واستعرضنا أهم تطبيقاتها الحالية في مجالات عدة، كما عرَّفنا بأهم الوظائف التي ينبغي الإلمام بها للمهتم في المجال وكيفية تطبيقها على النصوص للمساعدة في تجميع وتمثيل وتحليل النصوص المكتوبة والمنطوقة. وركَّزنا في هذا البحث على التعلم العميق في توليد اللغات الطبيعية. ففصلنا أولاً ماهية التعلم العميق وكيفية عمله، ثم عرضنا تجربتنا العملية، خطوة بخطوة، في توليد الشعر العربي باستخدام خوارزميات التعلم العميق.

إن اللغة العربية ثرية بالدلالات اللفظية والقواعد الصرفية والنحوية والتي تجعلها واحدة من أروع اللغات على الإطلاق. لذا، فلا بد من التشجيع على الخوض في مجال معالجة اللغات الطبيعية—وبخاصة للغة العربية—خصوصاً مع تطور خوارزميات التعلم العميق التي تسهل معالجة اللغات الطبيعية وتطبيقاتها.

المراجع

[1] How many languages are there in the world? Linguistic Society of America. [online] Available at: https://www.linguisticsociety.org/content/how-many-languages-are-there-world [Accessed 20 May 2019].

[2] A. M. Turing. "Computing machinery and intelligence". Mind. pp. 433-460. 1950.

[3] Saygin. A.P.. Cicekli. I. and Akman. V.. 2000. Turing test: 50 years later. Minds and machines. 10(4). pp.463-518.

[4] Translator. IBM. [Online]. Available at: http://www-03.ibm.com/ibm/history/exhib- its/701/701_translator.html. [Accessed 22 May 2019].

[5] Hutchins. J.. 2005. The first public demonstration of machine translation: the Georgetown-IBM system. 7th January 1954. Publicación electrónica en: http://www. hutchinsweb. me. uk/GUIBM-2005. pdf.

[6] Lehnert. W.. 1975. What makes SAM run? Script based techniques for question answering. In Theoretical Issues in Natural Language Processing: Supplement.

[7] McKeown. K.R.. 1980. Paraphrasing using given and new information in a question-answer system. Technical Reports (CIS). p.723.

[8] Karttunen. L.. 1969. Discourse referents. In INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS COLING 1969: Preprint No. 70.

[9] Rivers. W.M.. 1972. Speaking in many tongues: Essays in foreign-language teaching.

[10] Klein. S. and Simmons. R.F.. 1963. A computational approach to grammatical coding of English words. Journal of the ACM (JACM). 10(3). pp.334-347.

[11] Màrquez. L. and Rodríguez. H.. 1998. April. Part-of-speech tagging using decision trees. In European Conference on Machine Learning (pp. 25-36). Springer. Berlin. Heidelberg.

[12] Church. K.W.. 1989. May. A stochastic parts program and noun phrase parser for unrestricted text. In International Conference on Acoustics. Speech. and Signal Processing.(pp. 695-698). IEEE.

[13] DeRose. S.J.. 1988. Grammatical category disambiguation by statistical optimization. Computational linguistics. 14(1). pp.31-39.

[14] Das. D. and Martins. A.F.. 2007. A survey on automatic text summarization. Literature Survey for the Language and Statistics II course at CMU. 4(192-195). p.57.

[15] Chowdhury. G.G.. 2010. Introduction to modern information retrieval. Facet publishing.

[16] Chaovalit. P. and Zhou. L.. 2005. January. Movie review mining: A comparison between supervised and unsupervised classification approaches. In Proceedings of the 38th annual Hawaii international conference on system sciences (pp. 112c-112c). IEEE.

[17] Pang. B.. Lee. L. and Vaithyanathan. S.. 2002. July. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics.

[18] Kotsiantis. S.B.. Zaharakis. I. and Pintelas. P.. 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 160. pp.3-24.

[19] Khan. A.. Baharudin. B.. Lee. L.H. and Khan. K.. 2010. A review of machine learning algorithms for text-documents classification. Journal of advances in information technology. 1(1). pp.4-20.

[20] Goldberg. D.E. and Holland. J.H.. 1988. Genetic algorithms and machine learning. Machine learning. 3(2). pp.95-99.

[21] Badr. I.. Zbib. R. and Glass. J.. 2008. Segmentation for English-to-Arabic statistical machine translation. Proceedings of ACL-08: HLT. Short Papers. pp.153-156.

[22] Ghaffar. S.A.. Fakhr. M.W. and Sheraton. C.. 2011. English to Arabic statistical machine translation system improvements using preprocessing and Arabic morphology analysis. Recent Researches in Mathematical Methods in Electrical Engineering and Computer Science. pp.50-54.

[23] Badr. I.. Zbib. R. and Glass. J.. 2009. March. Syntactic phrase reordering for English-to-Arabic statistical machine translation. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 86-93). Association for Computational Linguistics.

[24] Al-Haj. H. and Lavie. A.. 2012. The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation. Machine translation. 26(1-2). pp.3-24.

[25] El Kholy. A. and Habash. N.. 2012. Orthographic and morphological processing for English–Arabic statistical machine translation. Machine Translation. 26(1-2). pp.25-45.

[26] Hasan. S.. El Isbihani. A. and Ney. H.. 2006. May. Creating a Large-Scale Arabic to French Statistical MachineTranslation System. In LREC (pp. 855-858).

[27] Schwenk. H. and Senellart. J.. 2009. Translation model adaptation for an Arabic/French news translation system by lightly-supervised training. In In MT Summit.

[28] Guidere. M.. 2002. Toward corpus-based machine translation for standard Arabic. Translation Journal. 6(1).

[29] Green. S.. Heer. J. and Manning. C.D.. 2013. April. The efficacy of human post-editing for language translation. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 439-448). ACM.

[30] Ehab. R.. Gadallah. M. and Amer. E.. 2019. English-Arabic Hybrid Machine Translation System using EBMT and Translation Memory. International Journal of Advanced Computer Science and Applications. 10(1). pp.195-203.

[31] Marie-Sainte. S.L.. Alalyani. N.. Alotaibi. S.. Ghouzali. S. and Abunadi. I.. 2019. Arabic natural language processing and machine learning-based systems. IEEE Access. 7. pp.7011-7020.

[32] Menacer. M.A.. Langlois. D.. Jouvet. D.. Fohr. D.. Mella. O. and Smaïli. K.. 2019. May. Machine Translation on a parallel Code-Switched Corpus. In Canadian Conference on Artificial Intelligence (pp. 426-432). Springer. Cham.

[33] Lodhi. H.. Saunders. C.. Shawe-Taylor. J.. Cristianini. N. and Watkins. C.. 2002. Text classification using string kernels. Journal of Machine Learning Research. 2(Feb). pp.419-444.

[34] Cavnar. W.B. and Trenkle. J.M.. 1994. April. N-gram-based text categorization. In Proceedings of SDAIR-94. 3rd annual symposium on document analysis and information retrieval(Vol. 161175).

[35] Joulin. A.. Grave. E.. Bojanowski. P. and Mikolov. T.. 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

[36] McCallum. A. and Nigam. K.. 1998. July. A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752. No. 1. pp. 41-48).

[37] Forman. G.. 2003. An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research. 3(Mar). pp.1289-1305.

[38] Zhang. X.. Zhao. J. and LeCun. Y.. 2015. Character-level convolutional networks for text classification. In Advances in neural information processing systems (pp. 649-657).

[39] Lai. S.. Xu. L.. Liu. K. and Zhao. J.. 2015. February. Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence.

[40] Conneau. A.. Schwenk. H.. Barrault. L. and Lecun. Y.. 2016. Very deep convolutional networks for text classification. arXiv preprint arXiv:1606.01781.

[41] Khreisat. L.. 2006. Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. DMIN. 2006. pp.78-82.

[42] Al-Harbi. S.. Almuhareb. A.. Al-Thubaity. A.. Khorsheed. M.S. and Al-Rajeh. A.. 2008. Automatic Arabic text classification.

[43] El-Halees. A.M.. 2007. Arabic text classification using maximum entropy. Arabic Text Classification Using Maximum Entropy. 15(1).

[44] Elarian. Y.. Ahmad. I.. Awaida. S.. Al-Khatib. W. and Zidouri. A.. 2015. Arabic ligatures: analysis and application in text recognition. In 13th International Conference on Document Analysis and Recognition (ICDAR) (pp. 896-900). IEEE.

[45] Elarian. Y.. Ahmad. I.. Awaida. S.. Al-Khatib. W.G. and Zidouri. A.. 2015. An Arabic handwriting synthesis system. Pattern Recognition. 48(3). pp.849-861.

[46] Kanaan. G.. Al-Shalabi. R.. Ghwanmeh. S. and Al-Ma'adeed. H.. 2009. A comparison of text-classification techniques applied to Arabic text. Journal of the American society for information science and technology. 60(9). pp.1836-1844.

[47] Wong. K.F.. Wu. M. and Li. W.. 2008. August. Extractive summarization using supervised and semi-supervised learning. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp. 985-992). Association for Computational Linguistics.

[48] Murray. G.. Renals. S. and Carletta. J.. 2005. Extractive summarization of meeting recordings.

[49] Paulus. R.. Xiong. C. and Socher. R.. 2017. A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.

[50] Ganesan. K.. Zhai. C. and Han. J.. 2010. August. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) (pp. 340-348).

[51] Oufaida. H.. Nouali. O. and Blache. P.. 2014. Multilingual Summarization Experiments on English. Arabic and French (Résumé Automatique Multilingue Expérimentations sur l'Anglais. l'Arabe et le Français)[in French]. Proceedings of TALN 2014 (Volume 2: Short Papers). 2. pp.543-549.

[52] Froud. H.. Lachkar. A. and Ouatik. S.A.. 2013. Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering. arXiv preprint arXiv:1302.1612.

[53] Douzidia. F.S. and Lapalme. G.. 2004. Lakhas. an Arabic summarization system. Proceedings of DUC2004.

[54] Al-Saleh. A.B. and Menai. M.E.B.. 2016. Automatic Arabic text summarization: a survey. Artificial Intelligence Review. 45(2). pp.203-234.

[55] Azmi. A. and Al-Thanyyan. S.. 2009. September. Ikhtasir—A user selected compression ratio Arabic text summarization system. In 2009 International Conference on Natural Language Processing and Knowledge Engineering (pp. 1-7). IEEE.

[56] Azmi. A.M. and Al-Thanyyan. S.. 2012. A text summarizer for Arabic. Computer Speech & Language. 26(4). pp.260-273.

[57] Wang. J.H.. Chung. E.S. and Jang. M.G.. Electronics and Telecommunications Research Institute. 2008. Semi-automatic construction method for knowledge base of encyclopedia question answering system. U.S. Patent 7.428.487.

[58] Soricut. R. and Brill. E.. 2006. Automatic question answering using the web: Beyond the factoid. Information Retrieval. 9(2). pp.191-206.

[59] Green. C.C.. 1969. The application of theorem proving to question-answering systems (No. CS-138). STANFORD UNIV CALIF DEPT OF COMPUTER SCIENCE.

[60] Hammo. B.. Abu-Salem. H. and Lytinen. S.. 2002. July. QARAB: A question answering system to support the Arabic language. In Proceedings of the ACL-02 workshop on Computational approaches to semitic languages (pp. 1-11). Association for Computational Linguistics.

[61] Bekhti. S.. Rehman. A.. Al-Harbi. M. and Saba. T.. 2011. AQUASYS: An Arabic Question-Answering System Based on Extensive Question Analysis and Answer Relevance Scoring. International Journal of Academic Research. 3(4).

[62] Abdelnasser. H.. Ragab. M.. Mohamed. R.. Mohamed. A.. Farouk. B.. El-Makky. N. and Torki. M.. 2014. Al-Bayan: an Arabic question answering system for the Holy Quran. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP) (pp. 57-64).

[63] Abu-Jbara. A.. King. B.. Diab. M. and Radev. D.. 2013. Identifying opinion subgroups in arabic online discussions. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2. pp. 829-835).

[64] Abdul-Mageed. M.. Diab. M. and Kübler. S.. 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language. 28(1). pp.20-37.

[65] Al-Smadi. M.. Al-Ayyoub. M.. Jararweh. Y. and Qawasmeh. O.. 2019. Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological. syntactic and semantic features. Information Processing & Management. 56(2). pp.308-319.

[66] Elhadad. M.K.. Li. K.F. and Gebali. F.. 2019. March. Sentiment Analysis of Arabic and English Tweets. In Workshops of the International Conference on Advanced Information Networking and Applications (pp. 334-348). Springer. Cham.

[67] Al-Ayyoub. M.. Khamaiseh. A.A.. Jararweh. Y. and Al-Kabi. M.N.. 2019. A comprehensive survey of Arabic sentiment analysis. Information Processing & Management. 56(2). pp.320-342.

[68] McKeown. K.R.. 1982. June. The TEXT system for natural language generation: An overview. In Proceedings of the 20th annual meeting on Association for Computational Linguistics(pp. 113-120). Association for Computational Linguistics.

[69] Mann. W.C.. 1983. June. An overview of the Nigel text generation grammar. In Proceedings of the 21st annual meeting on Association for Computational Linguistics (pp. 79-84). Association for Computational Linguistics.

[70] Yan. F. and Mikolajczyk. K.. 2015. Deep correlation for matching images and text. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3441-3450).

[71] Tokui. S.. Oono. K.. Hido. S. and Clayton. J.. 2015. December. Chainer: a next-generation open source framework for deep learning. In Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS) (Vol. 5. pp. 1-6).

[72] Li. J.. Monroe. W.. Ritter. A.. Galley. M.. Gao. J. and Jurafsky. D.. 2016. Deep reinforcement learning for dialogue generation. arXiv preprint arXiv:1606.01541.

[73] Young. T.. Hazarika. D.. Poria. S. and Cambria. E.. 2018. Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine. 13(3). pp.55-75.

[74] Zhu. Y.. Lu. S.. Zheng. L.. Guo. J.. Zhang. W.. Wang. J. and Yu. Y.. 2018. June. Texygen: A benchmarking platform for text generation models. In The 41st International ACM SIGIR

Conference on Research & Development in Information Retrieval (pp. 1097-1100). ACM.

[75] Kaiser. L.M. and Vinyals. O.. Google LLC. 2019. Generating parse trees of text segments using neural networks. U.S. Patent Application 10/268.671.

[76] Lippi. M.. Montemurro. M.A.. Degli Esposti. M. and Cristadoro. G.. 2019. Natural Language Statistical Features of LSTM-Generated Texts. IEEE Transactions on Neural Networks and Learning Systems.

[77] Guo. J.. Lu. S.. Cai. H.. Zhang. W.. Yu. Y. and Wang. J.. 2018. April. Long text generation via adversarial training with leaked information. In Thirty-Second AAAI Conference on Artificial Intelligence.

[78] Souri. A.. El Maazouzi. Z.. Al Achhab. M. and El Mohajir. B.E.. 2018. April. Arabic Text Generation Using Recurrent Neural Networks. In International Conference on Big Data. Cloud and Applications (pp. 523-533). Springer. Cham.

[79] Yousef Elarian (Editor). "الحرف العربي والتقنية" (Arabic and Technology). 2015. King Abdullah International Center for Arabic Language (KAICAL). Riyadh. Saudi Arabia.

[80] Elarian. Y.. Idris. F.. 2011. A Lexicon of Connected Components for Arabic Optical Text Recognition. In First International Workshop on Frontiers in Arabic Handwriting Recognition. Istanbul. Turkey.

[81] Taji. D.. Khalifa. S.. Obeid. O.. Eryani. F. and Habash. N.. 2018. October. An Arabic Morphological Analyzer and Generator with Copious Features. In Proceedings of the Fifteenth Workshop on Computational Research in Phonetics. Phonology. and Morphology (pp. 140-150).

[82] Ibrahim. W. and Hardie. A.. 2018. Accessible Corpus Annotation for Arabic. Arabic Corpus Linguistics. p.56.

[83] Hull. D.A.. 1996. Stemming algorithms: A case study for detailed evaluation. Journal of the American Society for Information Science. 47(1). pp.70-84.

[84] Paice. C.D.. 1994. An evaluation method for stemming algorithms. In SIGIR'94 (pp. 42-50). Springer. London.

[85] Willett. P.. 2006. The Porter stemming algorithm: then and now. Program. 40(3). pp.219-223.

[86] Hull. D.A. and Grefenstette. G.. 1996. A detailed analysis of English stemming algorithms. In Xerox Research and Technology.

[87] Taghva. K.. Elkhoury. R. and Coombs. J.. 2005. April. Arabic stemming without a root dictionary. In International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II (Vol. 1. pp. 152-157). IEEE.

[88] Hadni. M.. Ouatik. S.A. and Lachkar. A.. 2013. Effective Arabic stemmer based hybrid approach for Arabic text categorization. International Journal of Data Mining & Knowledge Management Process. 3(4). p.1.

[89] Al-Kabi. M.N.. Kazakzeh. S.A.. Ata. B.M.A.. Al-Rababah. S.A and Alsmadi. I.M.. 2015. A novel root based Arabic stemmer. Journal of King Saud University-Computer and Information Sciences. 27(2). pp.94-103.

[90] Vergyri. D. and Kirchhoff. K.. 2004. August. Automatic diacritization of Arabic for acoustic modeling in speech recognition. In Proceedings of the workshop on computational approaches to Arabic script-based languages (pp. 66-73). Association for Computational Linguistics.

[91] Fadel. A.. Tuffaha. I.. Al-Jawarneh. B. and Al-Ayyoub. M.. 2019. Arabic Text Diacritization Using Deep Neural Networks. arXiv preprint arXiv:1905.01965.

[92] Punyakanok. V.. Roth. D. and Yih. W.T.. 2008. The importance of syntactic parsing and inference in semantic role labeling. Computational Linguistics. 34(2). pp.257-287.

[93] Chiang. D.. Diab. M.. Habash. N.. Rambow. O. and Shareef. S.. 2006. Parsing arabic dialects. In 11th Conference of the European Chapter of the Association for Computational Linguistics.

[94] McDonald. R.. Pereira. F.. Ribarov. K. and Hajič. J.. 2005. October. Non-projective dependency parsing using spanning tree algorithms. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 523-530). Association for Computational Linguistics.

[95] Nadeau. D. and Sekine. S.. 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes. 30(1). pp.3-26.

[96] LeCun. Y.. Bengio. Y. and Hinton. G.. 2015. Deep learning. nature. 521(7553). p.436.

[97] Goodfellow. I.. Bengio. Y. and Courville. A.. 2016. Deep learning. MIT press.

[98] Mikolov. T.. Karafiát. M.. Burget. L.. Černocký. J. and Khudanpur. S.. 2010. Recurrent neural network based language model. In Eleventh annual conference of the international speech communication association.

[99] Mikolov. T.. Chen. K.. Corrado. G. and Dean. J.. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[100] Arabic Poet. [Online]. Available at: https://twitter.com/AI_Sha3er. [Accessed 25 May 2019].

[101] Generate Arabic Poems. [Online]. Available at: https://github.com/Gharibw/Char-RNN-Arabic. [Accessed 25 May 2019].

[102] Tweets and Hashtag Harvester using Python. [Online]. Available at: https://github.com/Gharibw/Tweets_Harvester. [Accessed 25 May 2019].

فهرس الكتاب

مباحث لغوية ١٠

تطبيقات الذكاء الاصطناعي في خـــدمة اللغـــة العربية

يُصدِر مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية هذا الكتاب ضمن سلسلة (مباحث لغوية)، وذلك وفق خطة عمل مقسمة إلى مراحل، لموضوعات علمية رأى المركز حاجة المكتبة اللغوية العربية إليها، أو إلى بدء النشاط البحثي فيها، واجتهد في استكتاب نخبة من المحررين والمؤلفين للنهوض بعنوانات هذه السلسلة على أكمل وجه.

ويهدف المركز من وراء ذلك إلى تنشيط العمل في المجالات التي تُنَبّه إليها هذه السلسلة، سواء أكان العمل علميا بحثيا، أم عمليا تنفيذيا، ويدعو المركز الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسلة.

وتودّ الأمانة العامة أن تشيد بجهد السادة المؤلفين، وجهد محرر الكتاب، على ما تفضلوا به من رؤى وأفكار لخدمة العربية في هذا السياق البحثي.

والشكر والتقدير الوافر لمعالي وزير التعليم المشرف العام على المركز، الذي يحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة.

والدعوة موجّهة إلى جميع المختصين والمهتمين للتواصل مع المركز؛ لبناء المشروعات العلمية، وتكثيف الجهود، والتكامل نحو تمكين لغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

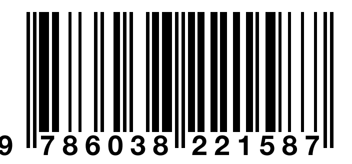
الأمين العام للمركز

د. عبدالله بن صالح الوشمي



ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف: ٠٠٩٦٦١١٢٥٨٧٢٦٨ – ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa




9 786038 221587